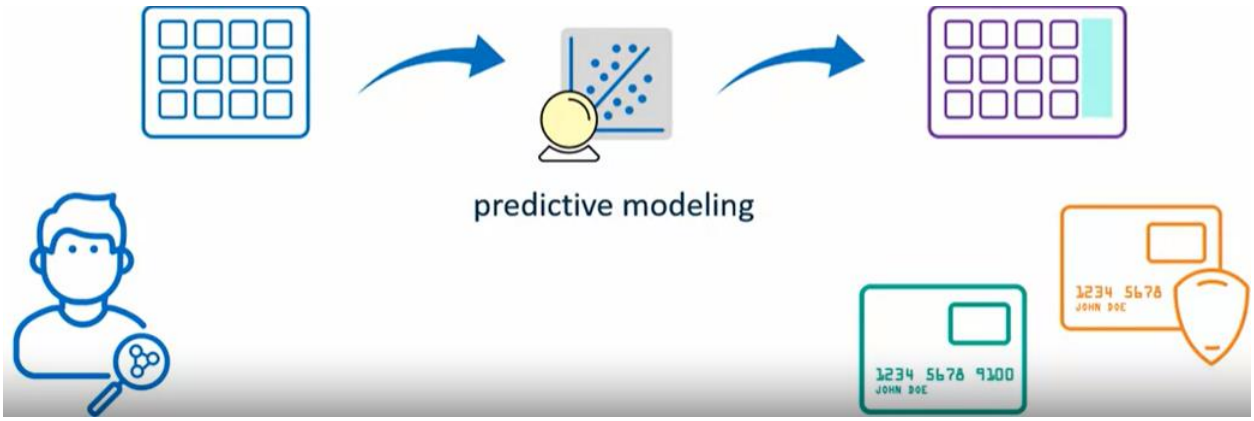
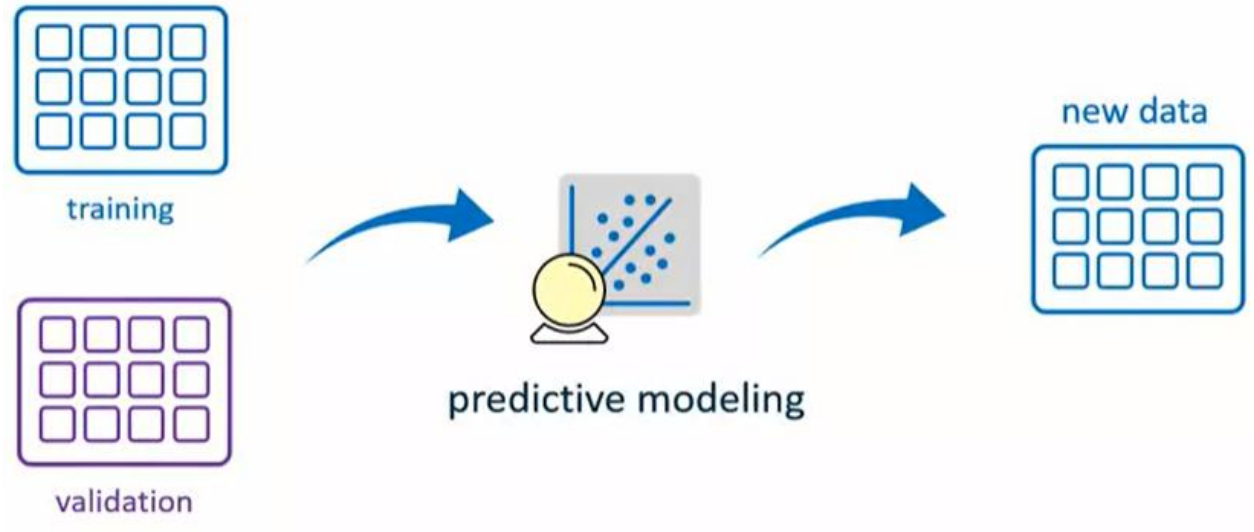
SBA Statistical Business Analyst with SAS

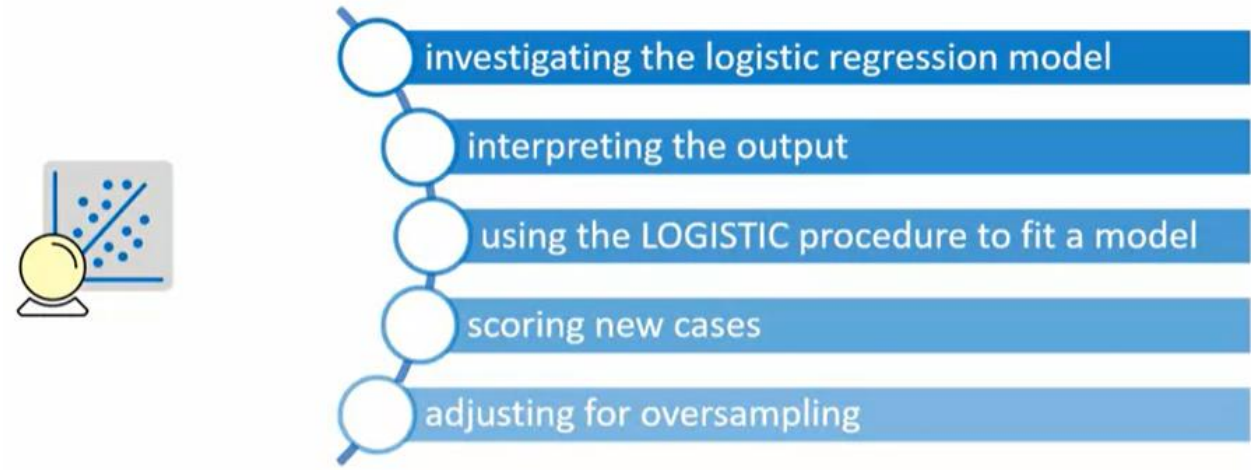
SBA3 Predictive Modeling with Logistic Regression

W2 Understanding the Logistic Regression Model and Correcting for Oversampling

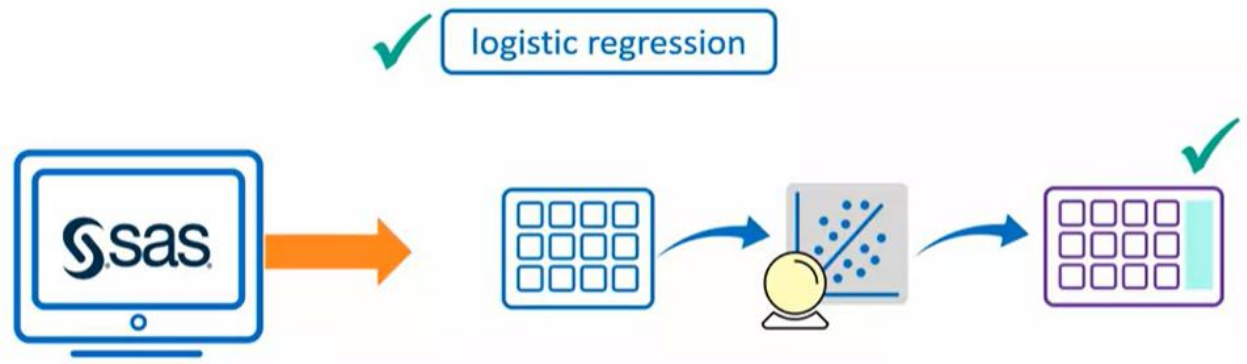
Overview







Introduction



In this topic, you learn to do the following:

- describe the mathematical representation of logistic regression
- describe several concepts that are important when you interpret logistic regression models
- fit a logistic regression model in the LOGISTIC procedure
- use PROC LOGISTIC to score new cases

Understanding the Logistic Regression Model

Logistic Regression

$$\text{logit}(p_i)=\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}$$

represents case *i*

Logistic Regression

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

✗ $E(y) =$

sum of the linear combination of predictors

$-\infty$ ⟷ ∞

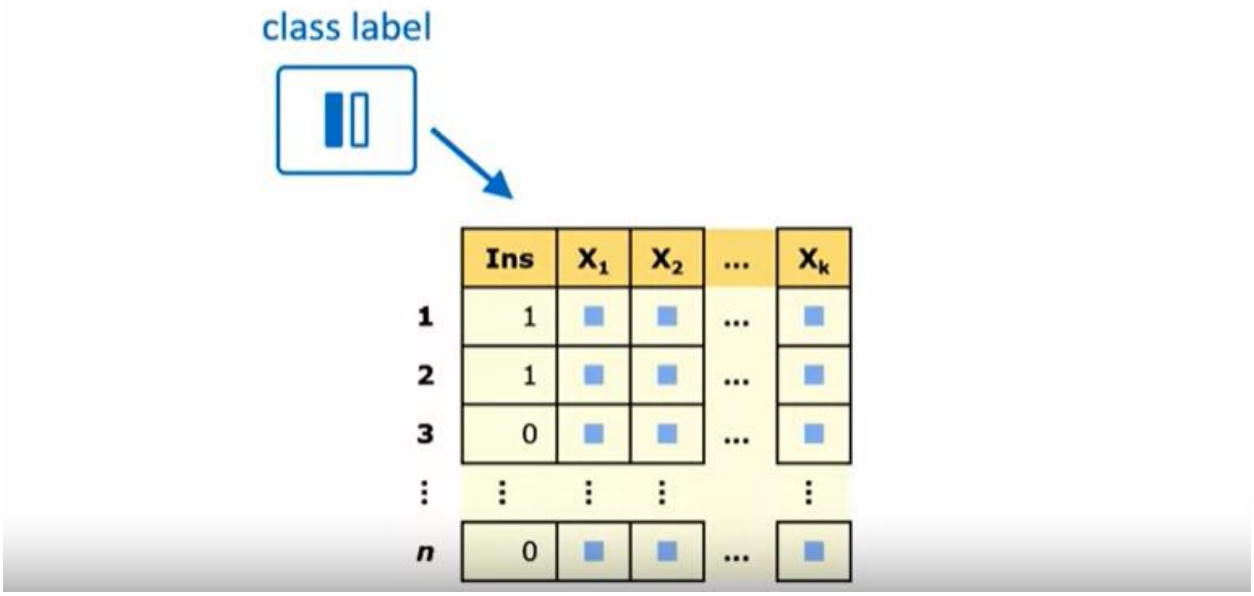
probability

0 — 1

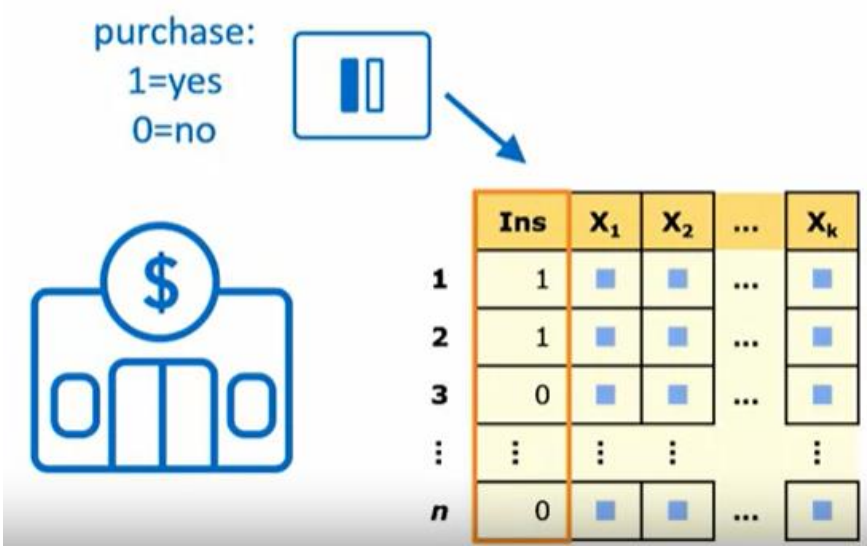
$$p = E(y|\underline{\mathbf{x}}) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}$$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0+\beta_1 x_1+\ldots+\beta_k x_k$$

$$\text{logit}(E(\text{Ins}|\underline{\mathbf{x}})) = \text{logit}(E(y|\underline{\mathbf{x}})) = \text{logit}(p) = \beta_0+\beta_1 x_1+\ldots+\beta_k x_k$$



$$\text{logit}(E(\text{Ins}|\underline{\mathbf{x}})) = \text{logit}(E(y|\underline{\mathbf{x}})) = \text{logit}(p) = \beta_0+\beta_1 x_1+\ldots+\beta_k x_k$$

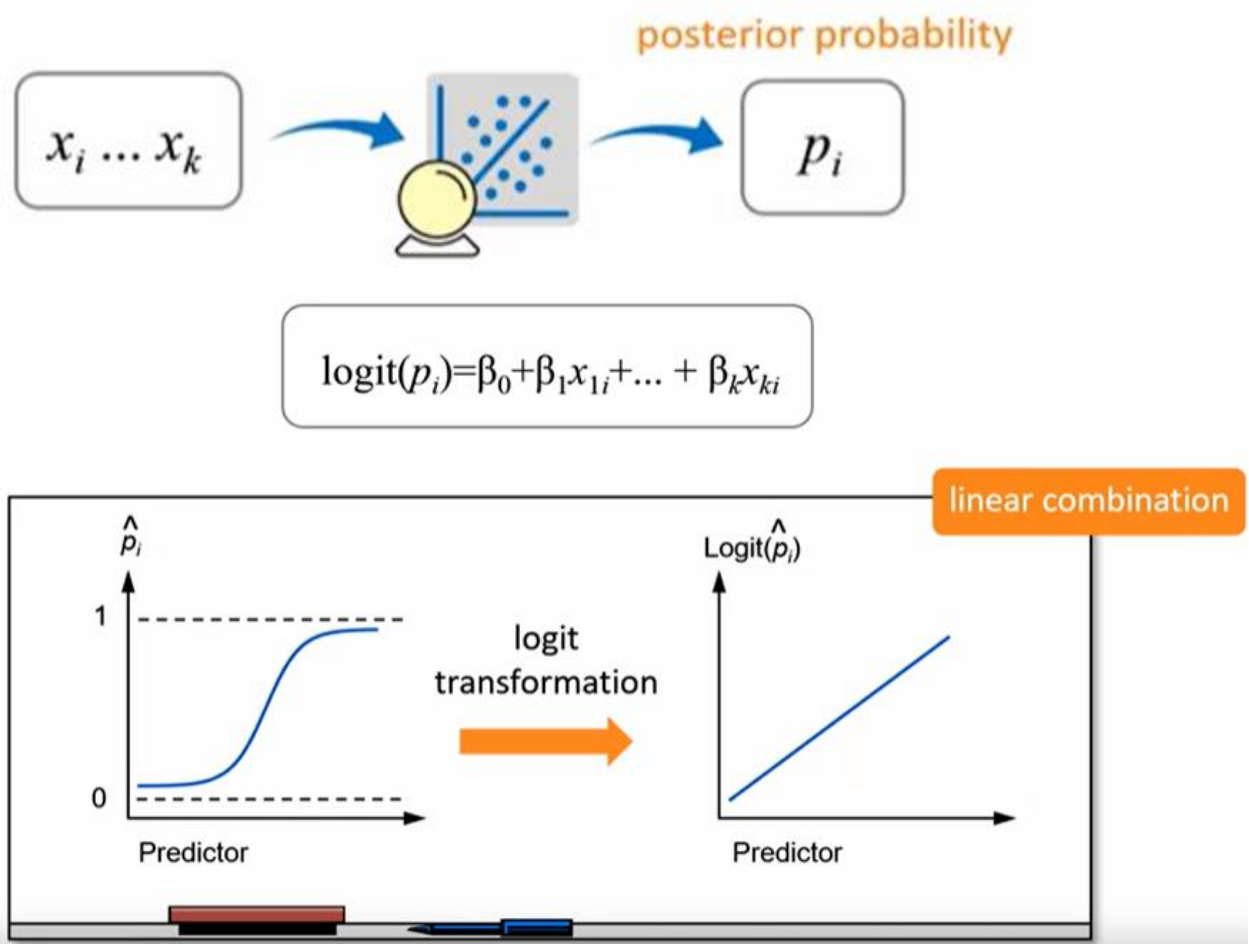


$$\text{logit}(p_i)=\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}$$

posterior probability

$$y_i = \text{Ins}_i \longleftarrow p_i = E(\text{Ins}_i|x_i)$$

probability that $y = 1$, given the inputs


posterior probability
$x_i \dots x_k$
p_i
$\text{logit}(p_i)=\beta_0+\beta_1 x_{1i}+\dots+\beta_k x_{ki}$
linear combination
$\hat{p}_i$
1
0
Predictor
logit
transformation
$\text{Logit}(\hat{p}_i)$
Predictor

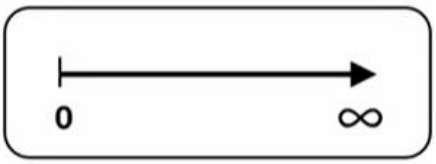
Constraining the Posterior Probability Using the Logit Transformation

Logit Transformation

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta_i$$

Transformations to find the logit(p_i)
odds transformation
log of the odds transformation

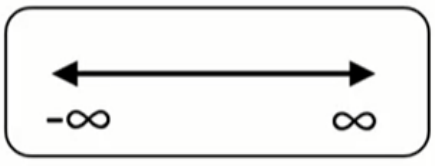
Transformations to find the logit(p_i)
odds transformation
log of the odds transformation



Logit Transformation

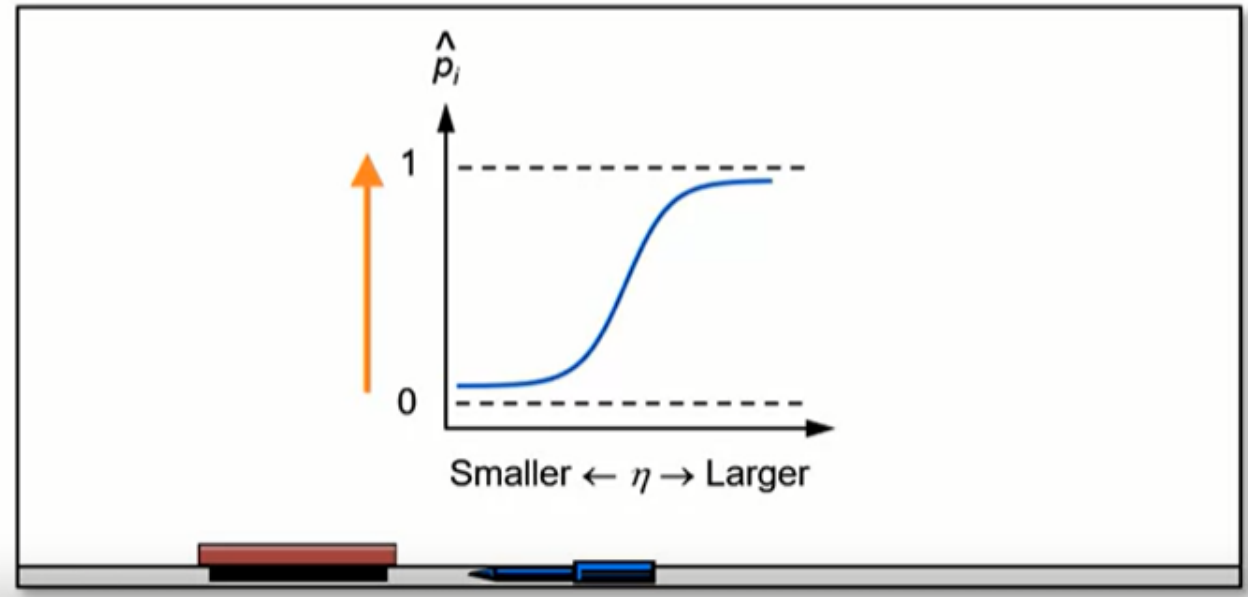
$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta_i$$

Transformations to find the logit(p_i)
odds transformation
log of the odds transformation



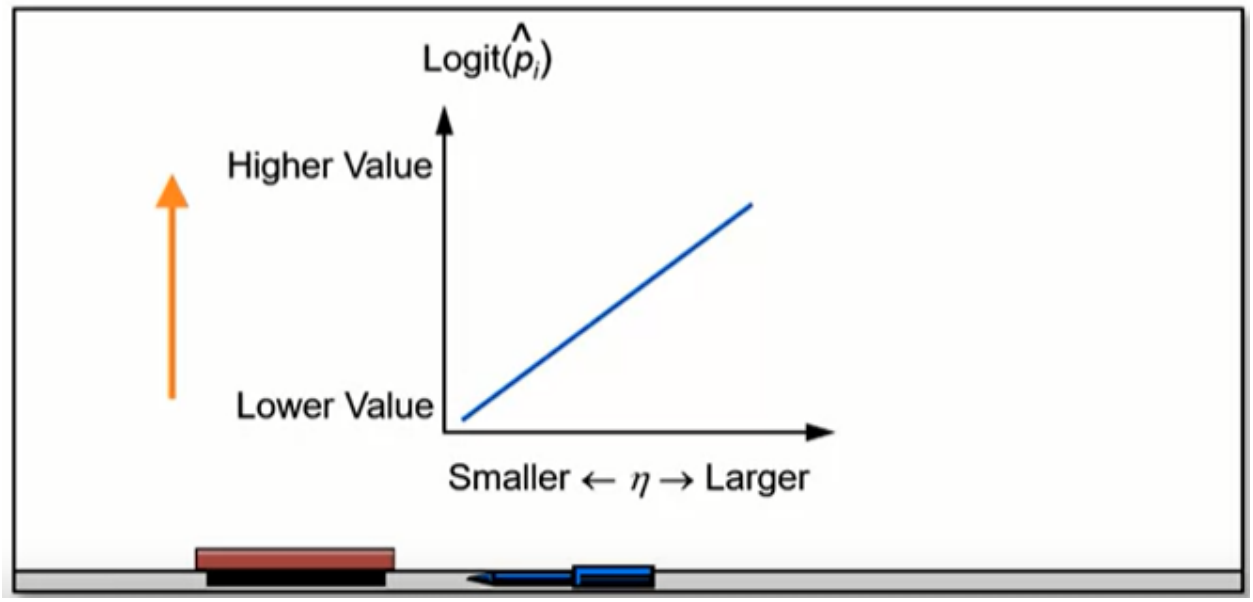
Logit Transformation

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta_i$$



Logit Transformation

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta_i$$

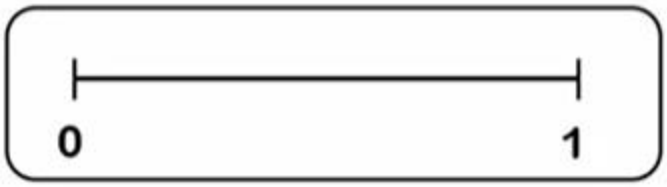


$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki})}} = \frac{1}{1+e^{-\eta_i}}$$

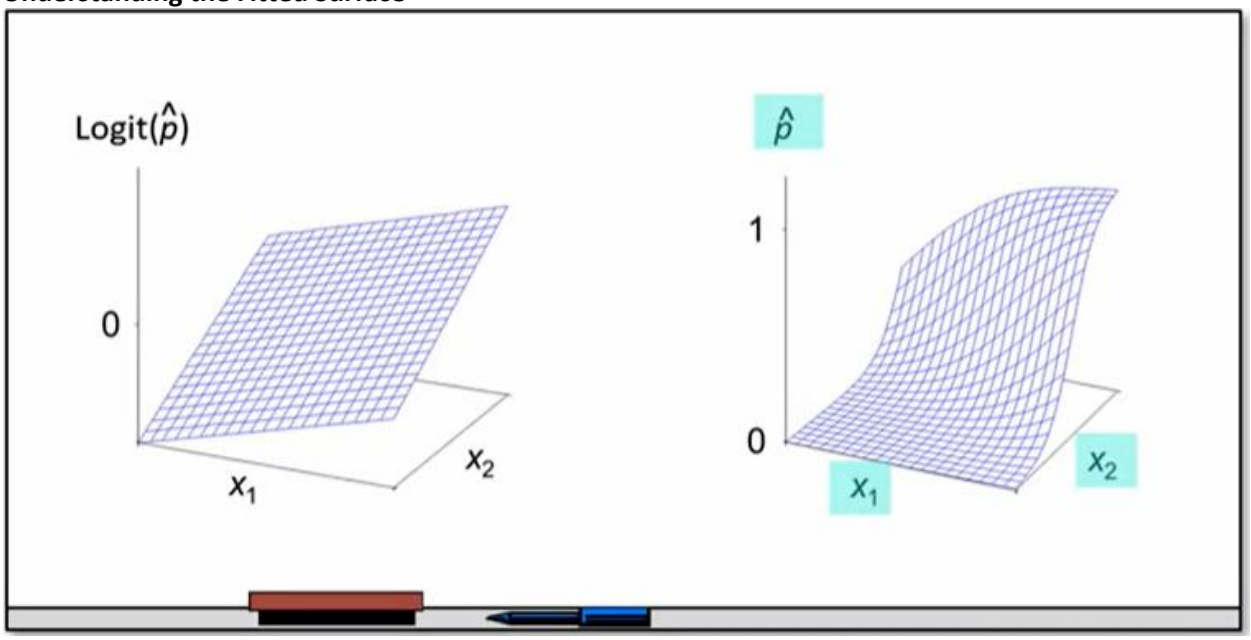
p_i

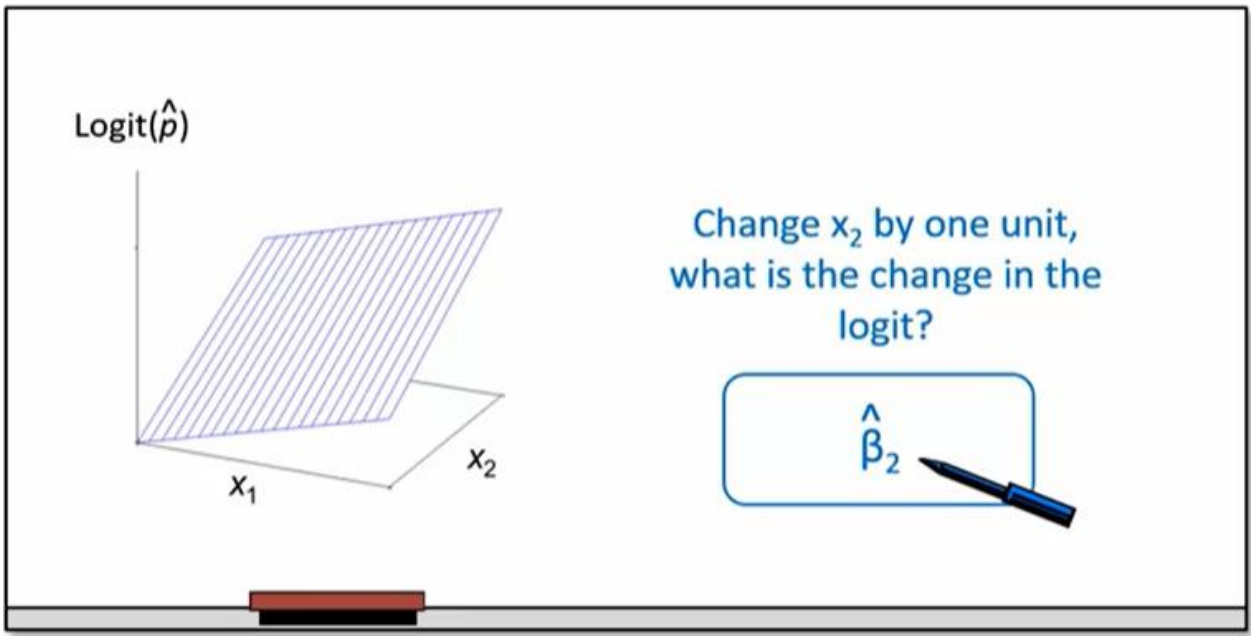
$$p_i = \frac{1}{1+e^{-\eta_i}}$$

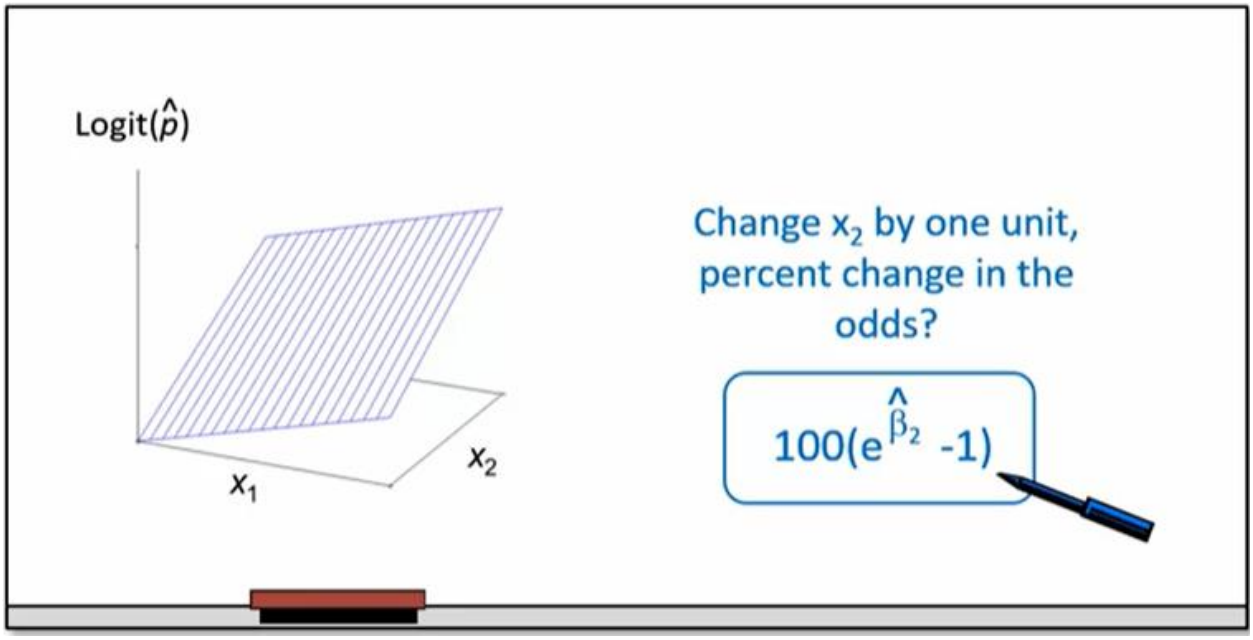
p_i

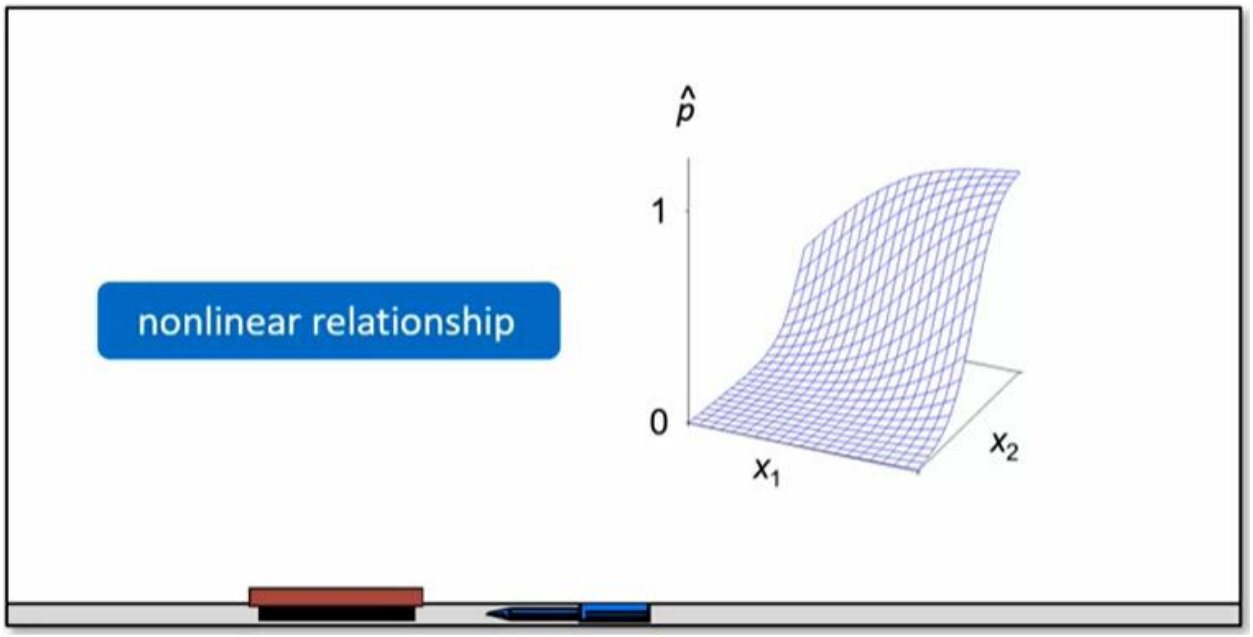


Understanding the Fitted Surface

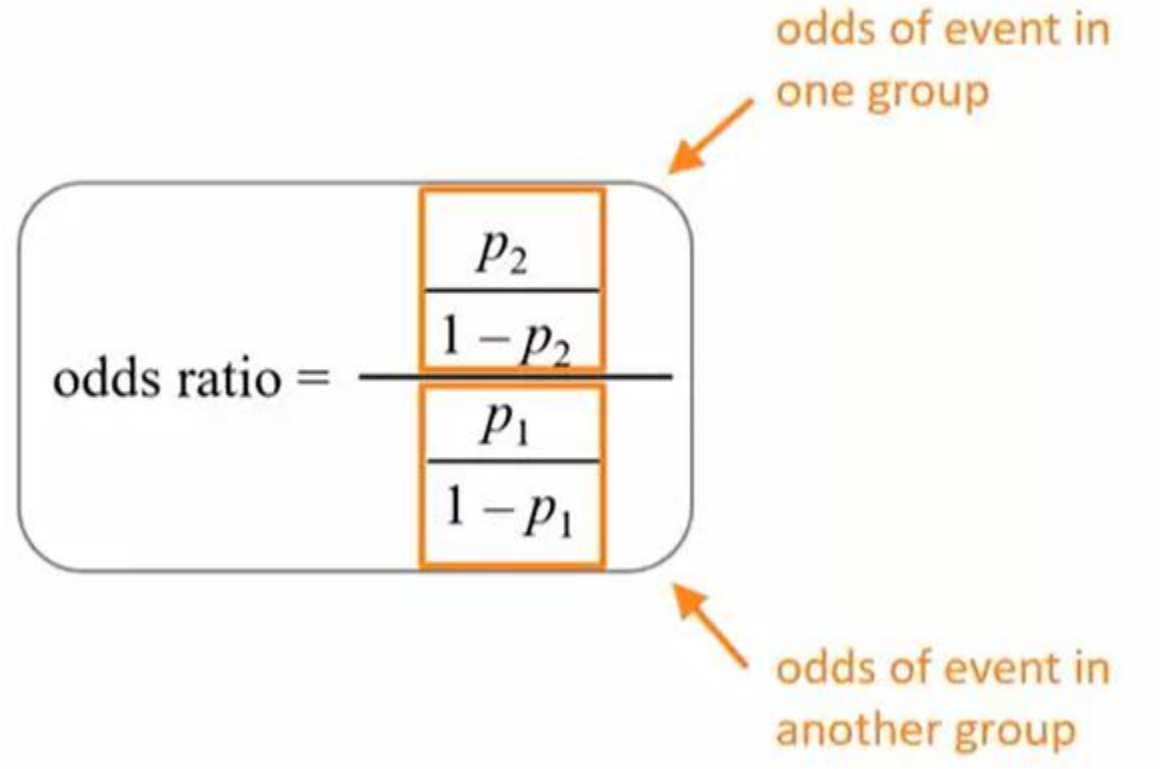





Logit($\hat{p}$)
x_1
x_2
Change x_2 by one unit, percent change in the odds?
$100(e^{\hat{\beta}_2} - 1)$


nonlinear relationship
$\hat{p}$
1
0
x_1
x_2

Interpreting the Model by Calculating the Odds Ratio

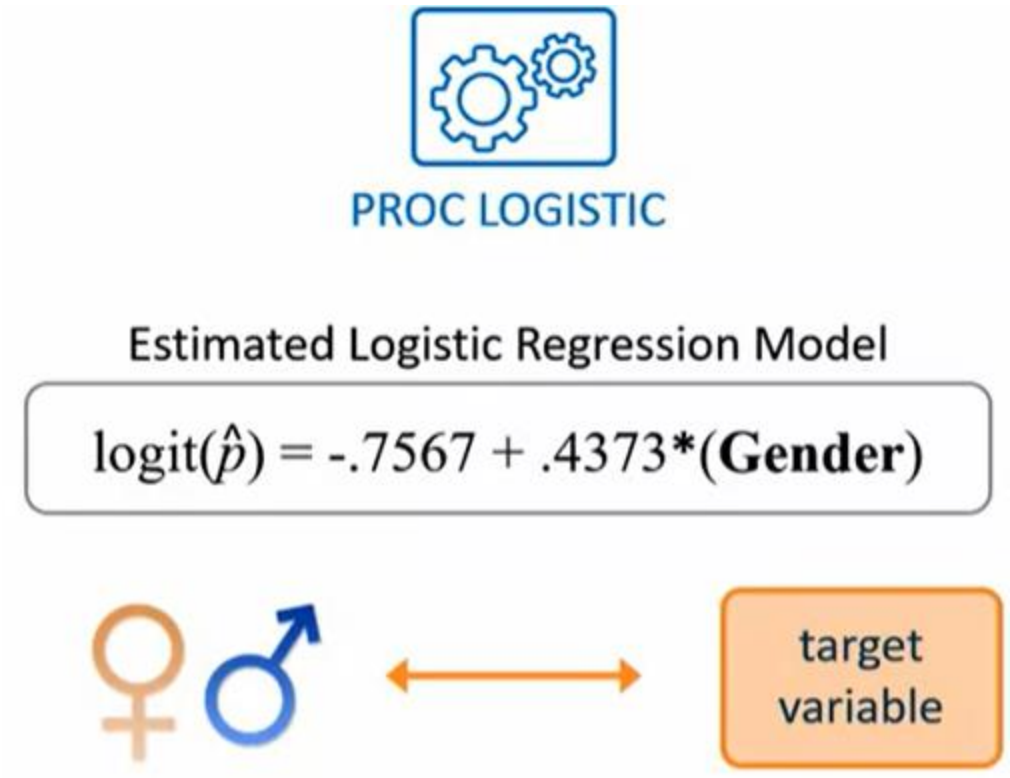


$$\frac{\text{odds of event for Group B}}{\text{odds of event for Group A}}$$

$\exp(\beta)$

odds ratios for a one-unit change in each variable

e^{β_1} e^{β_2} ... e^{β_k}


PROC LOGISTIC
Estimated Logistic Regression Model
$\text{logit}(\hat{p}) = -.7567 + .4373*(\textbf{Gender})$
target variable

PROC LOGISTIC

Estimated Logistic Regression Model

$$\text{logit}(\hat{p}) = -.7567 + .4373*(\textbf{Gender})$$

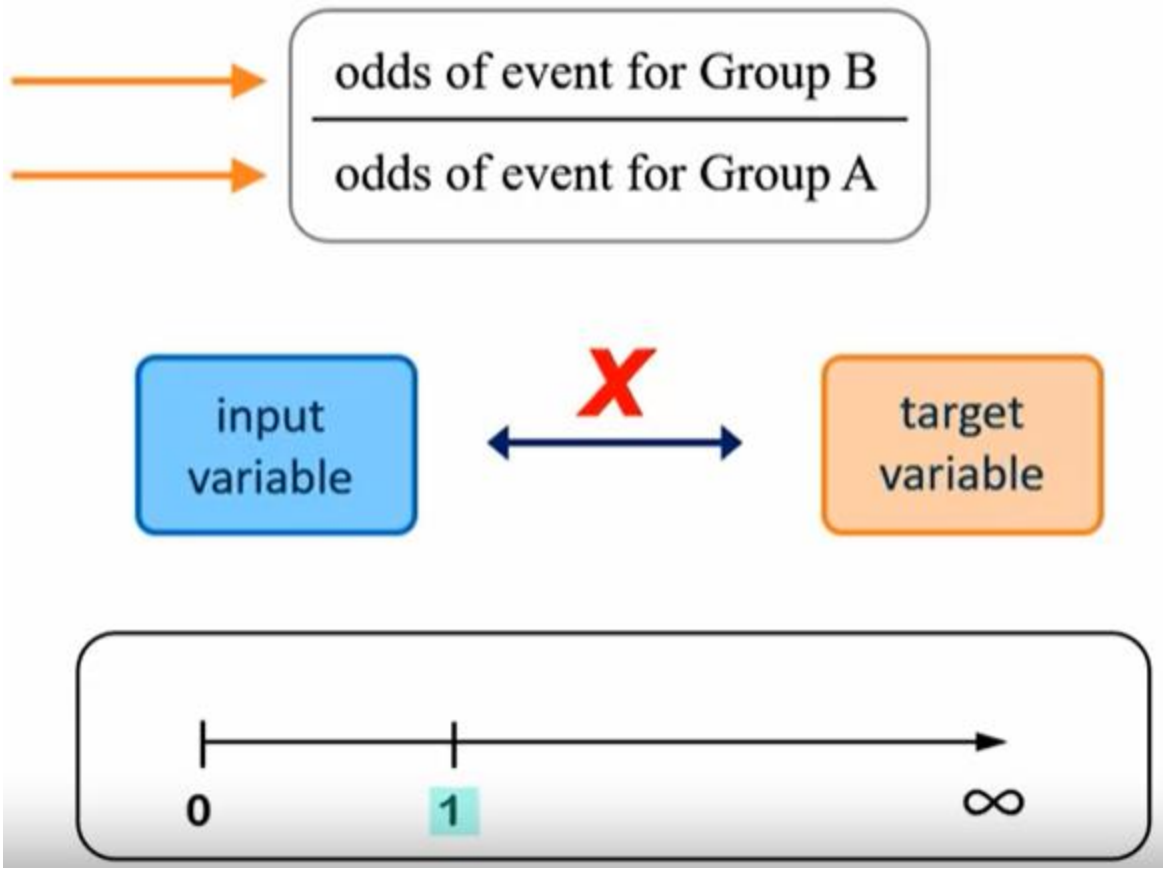
 = 1

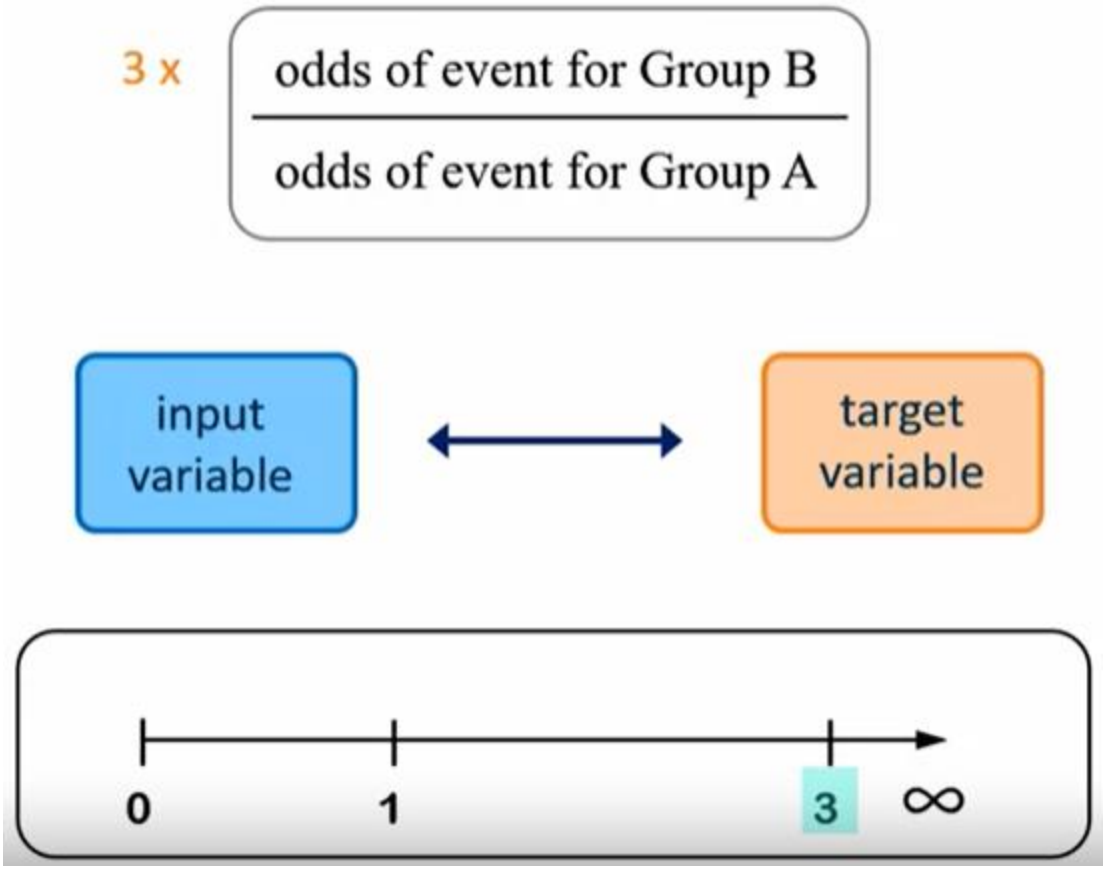
 = 0

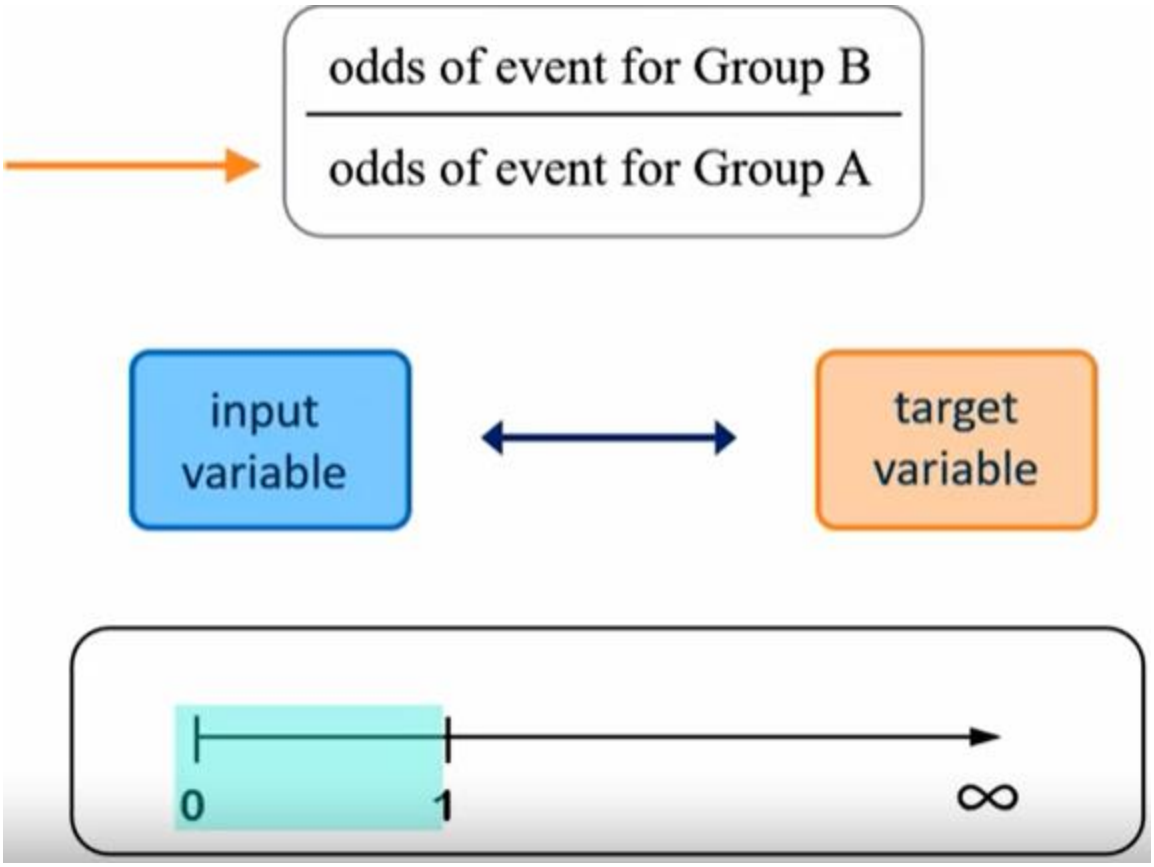
Estimated Odds Ratio for **Gender**

$$\text{odds ratio} = \frac{(e^{-.7567 + .4373})}{(e^{-.7567})} = e^{.4373} = 1.55$$

1.55 x $\dfrac{\text{odds that females have the outcome}}{\text{odds that males have the outcome}}$


odds of event for Group B
odds of event for Group A
input variable
X
target variable
0
1
∞


3 x
odds of event for Group B
odds of event for Group A
input variable
target variable
0
1
3
∞



Question 2.01

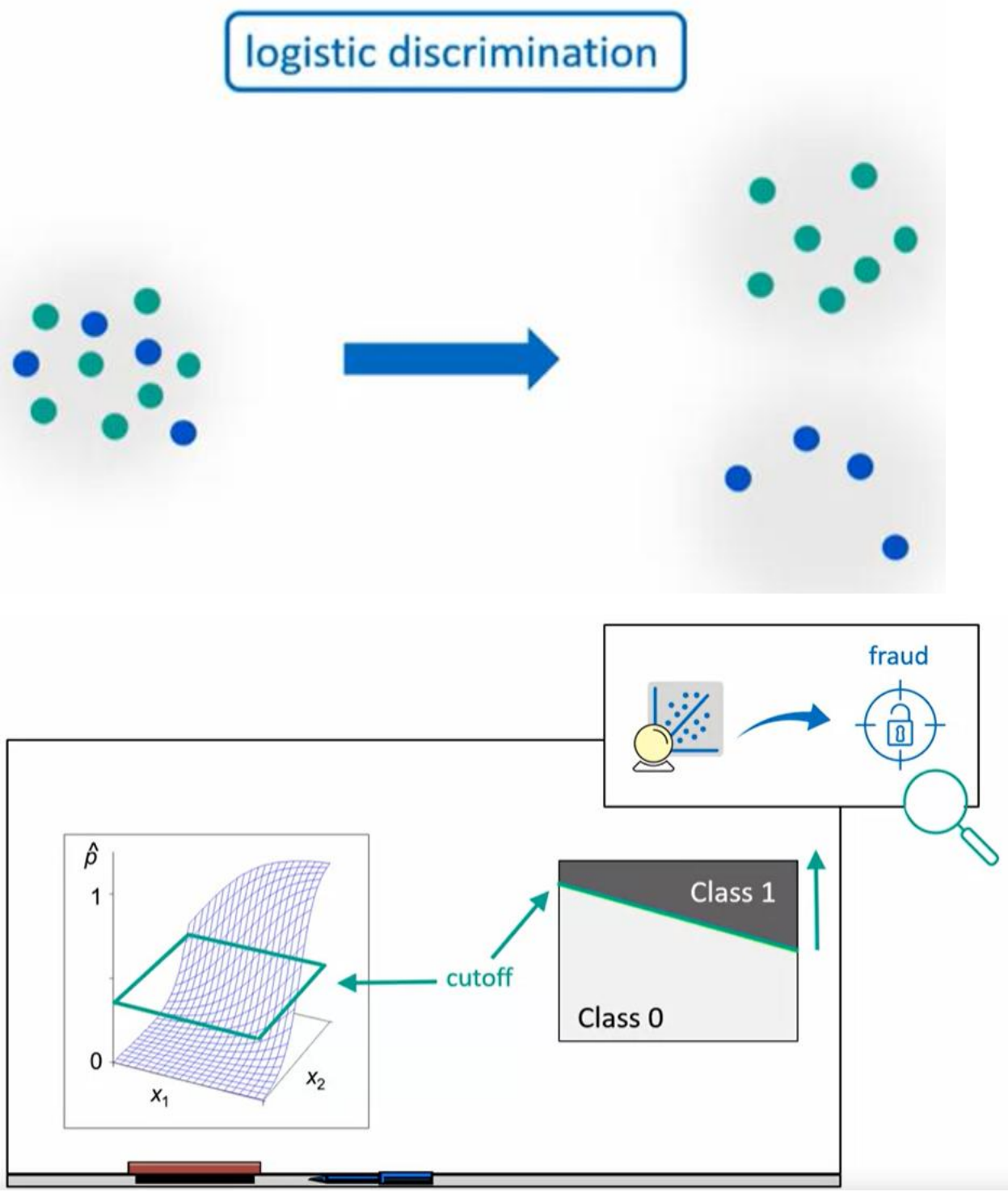
The odds ratio for a $1000 increase in income is 1.074. What does this mean for every $1000 increase in income?

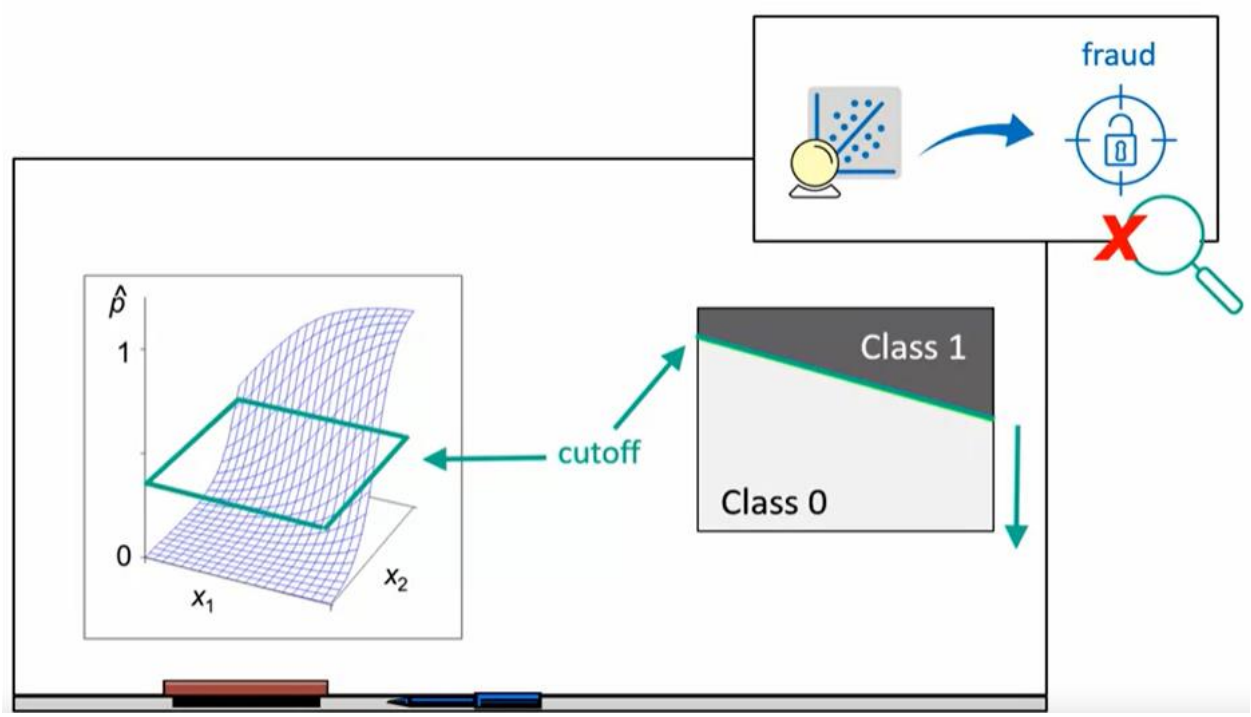
The odds of the event increase 7.4%.

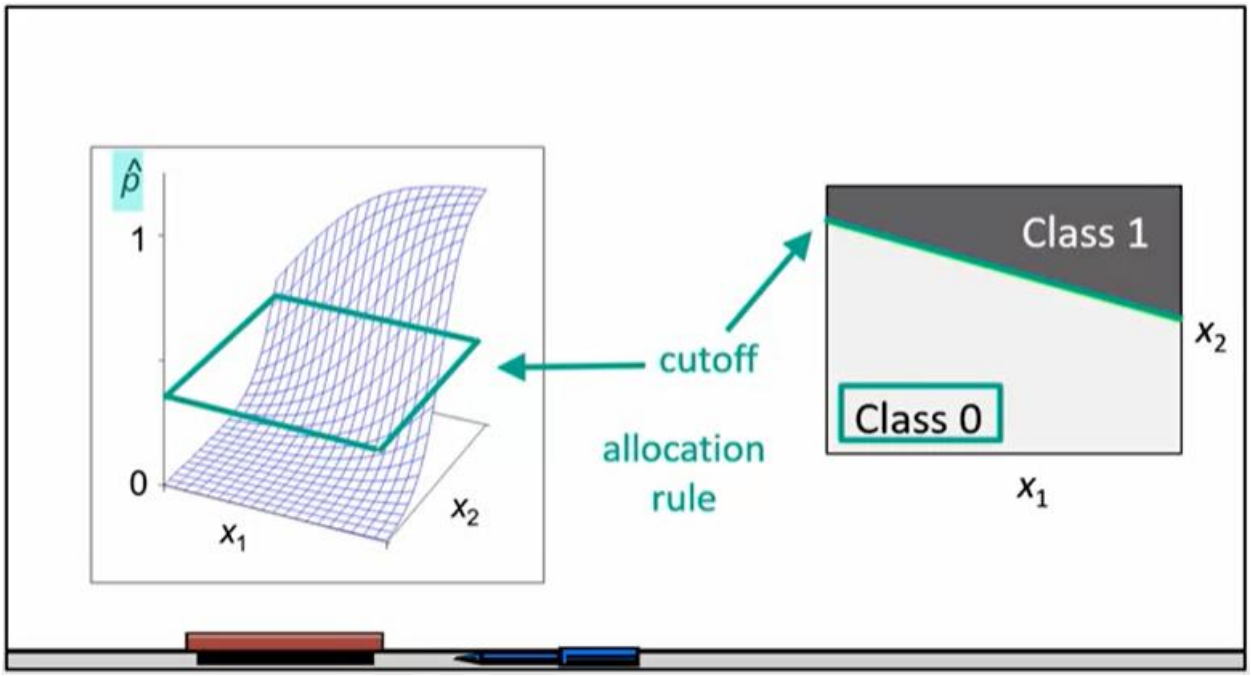
Correct

An odds ratio shows the change in the odds, not the change in the logit or probability. Logistic regression parameter estimates show the change in the logit.

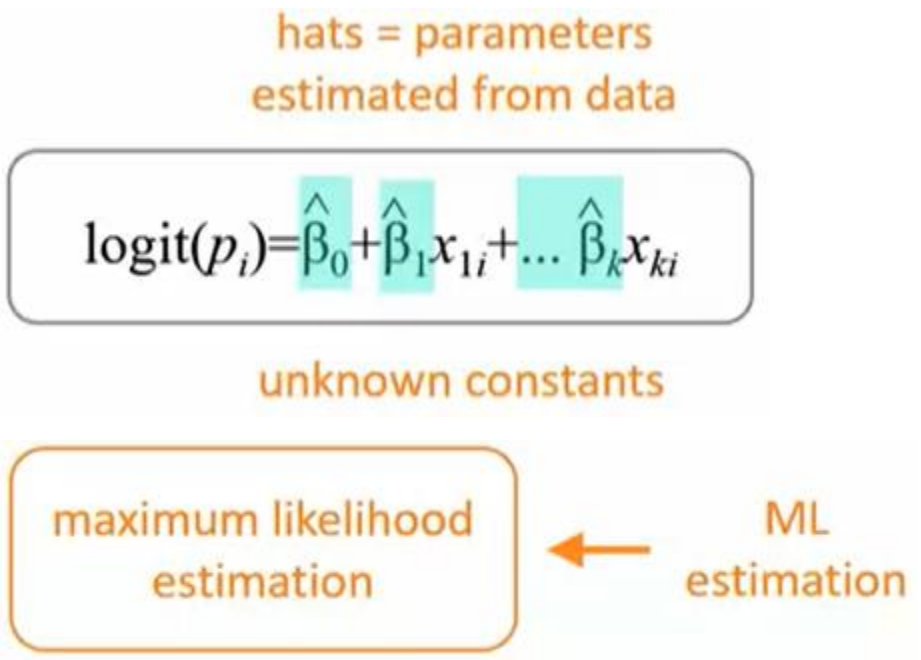
Understanding Logistic Discrimination




fraud
$\hat{p}$
1
0
x_1
x_2
cutoff
Class 1
Class 0


$\hat{p}$
1
0
x_1
x_2
cutoff
allocation
rule
Class 1
Class 0
x_2
x_1

Estimating Unknown Parameters Using Maximum Likelihood Estimation



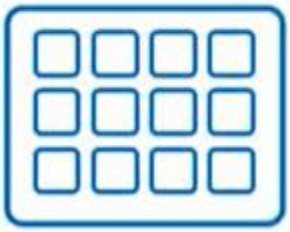
joint probability density function of the data

$$\text{logit}(p_i)=\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\ldots\ \hat{\beta}_k x_{ki}$$

probability that the observed data would occur

$$\text{logit}(p_i)=\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\ldots\ \hat{\beta}_k x_{ki}$$

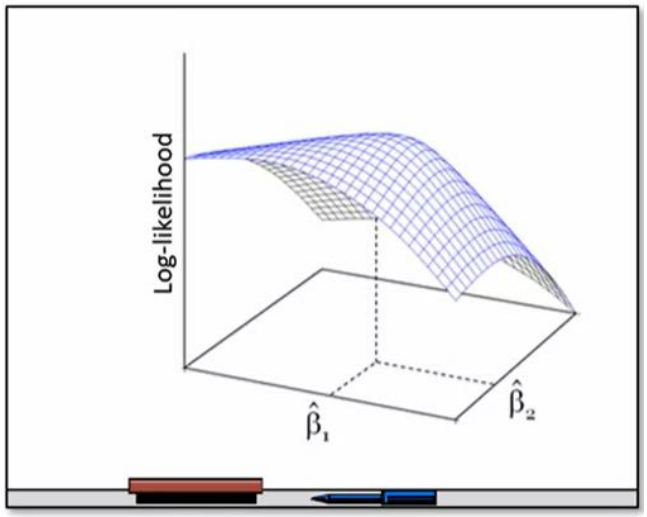
probability of obtaining the sample data



$$\text{logit}(p_i)=\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\ldots\ \hat{\beta}_k x_{ki}$$

log of the likelihood

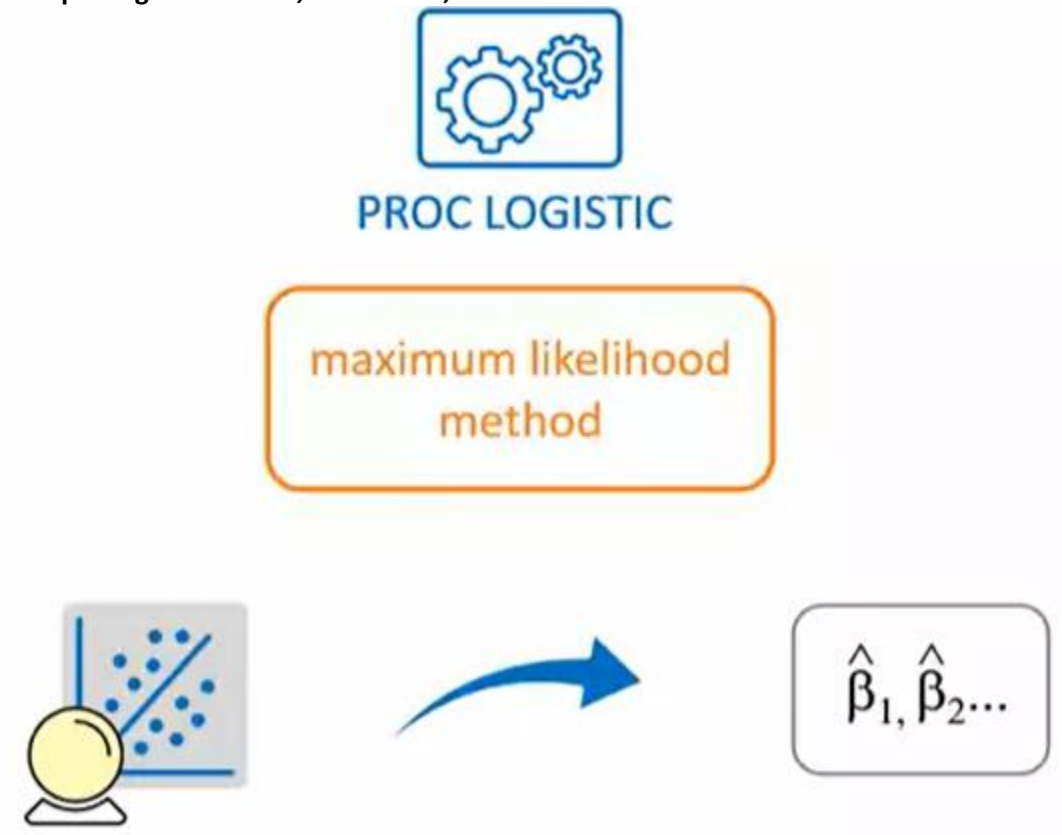
Maximum Likelihood Estimation
1. Use an iterative process to find the surface.
2. Report the parameters that correspond to the highest point on the surface.

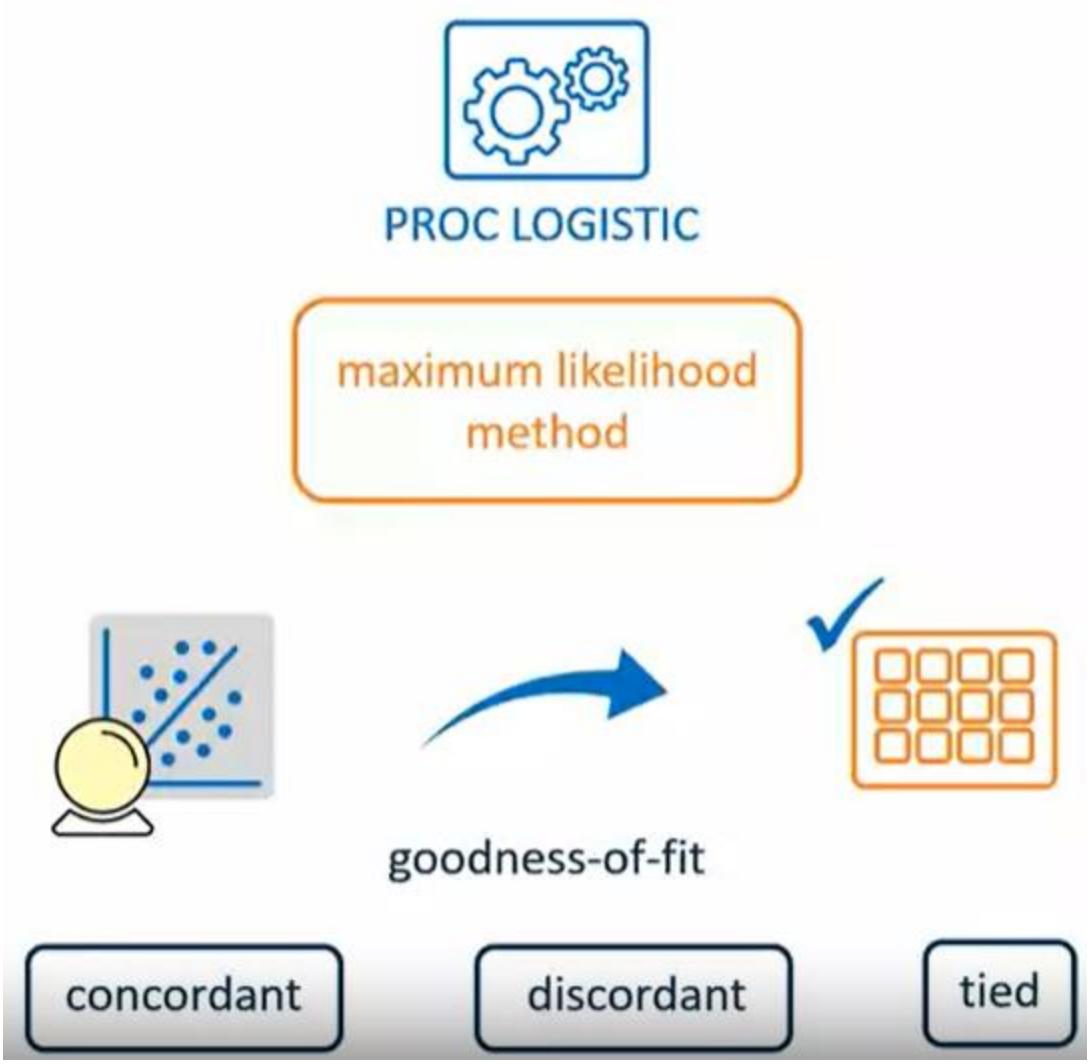


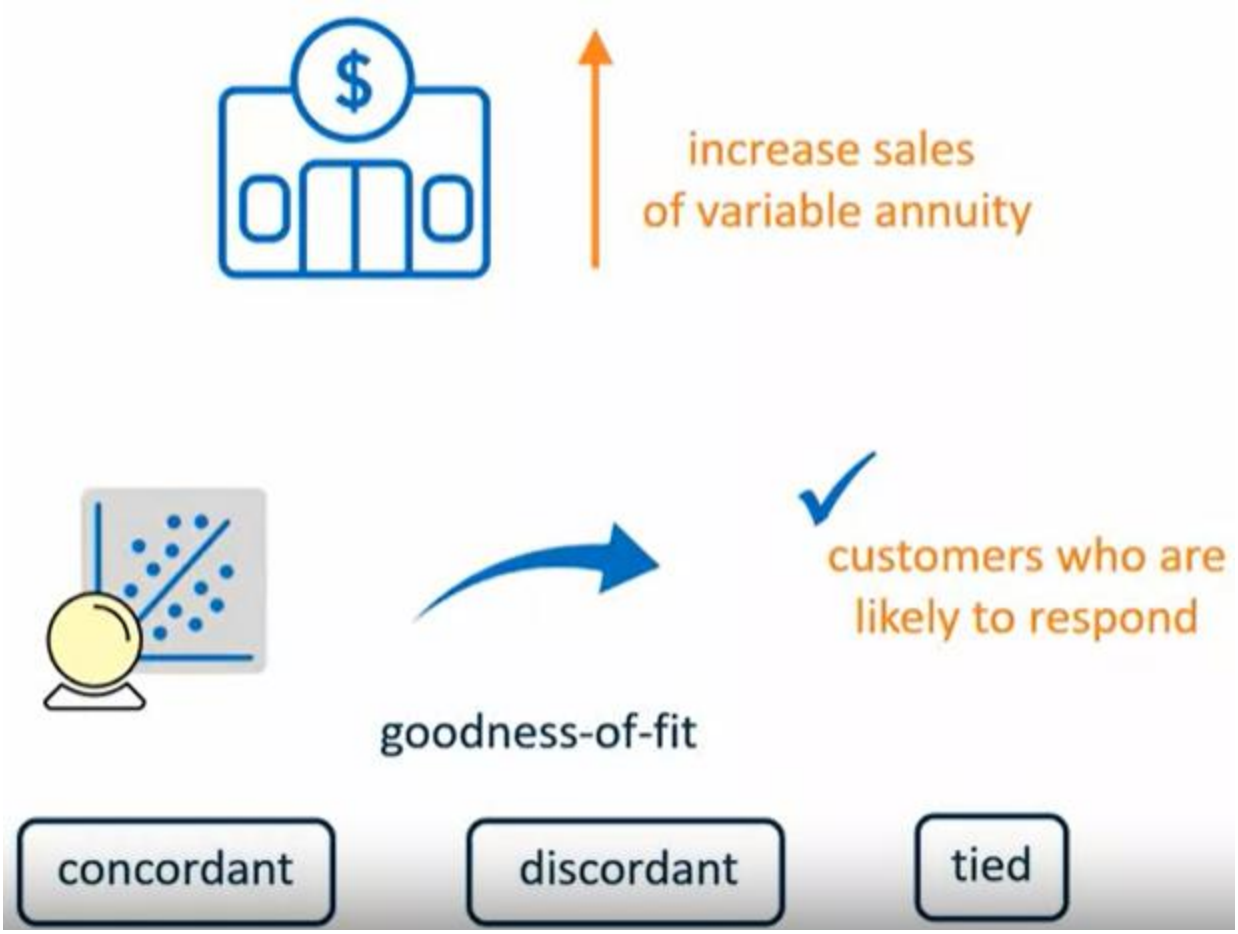
determine the combination of parameter values that maximizes the likelihood

Ordinary Least Squares Regression	Logistic Regression
closed form solution	iterative optimization algorithm

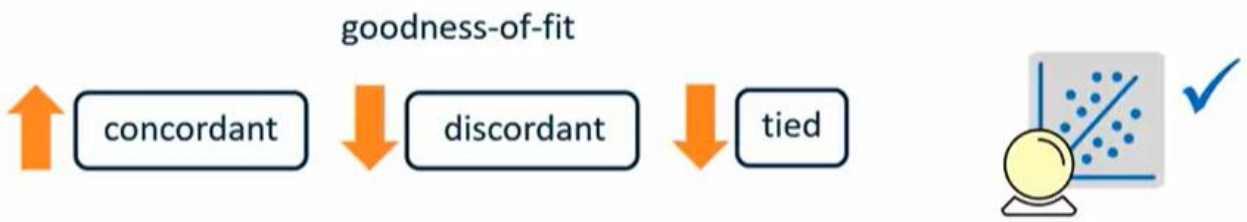
Interpreting Concordant, Discordant, and Tied Pairs



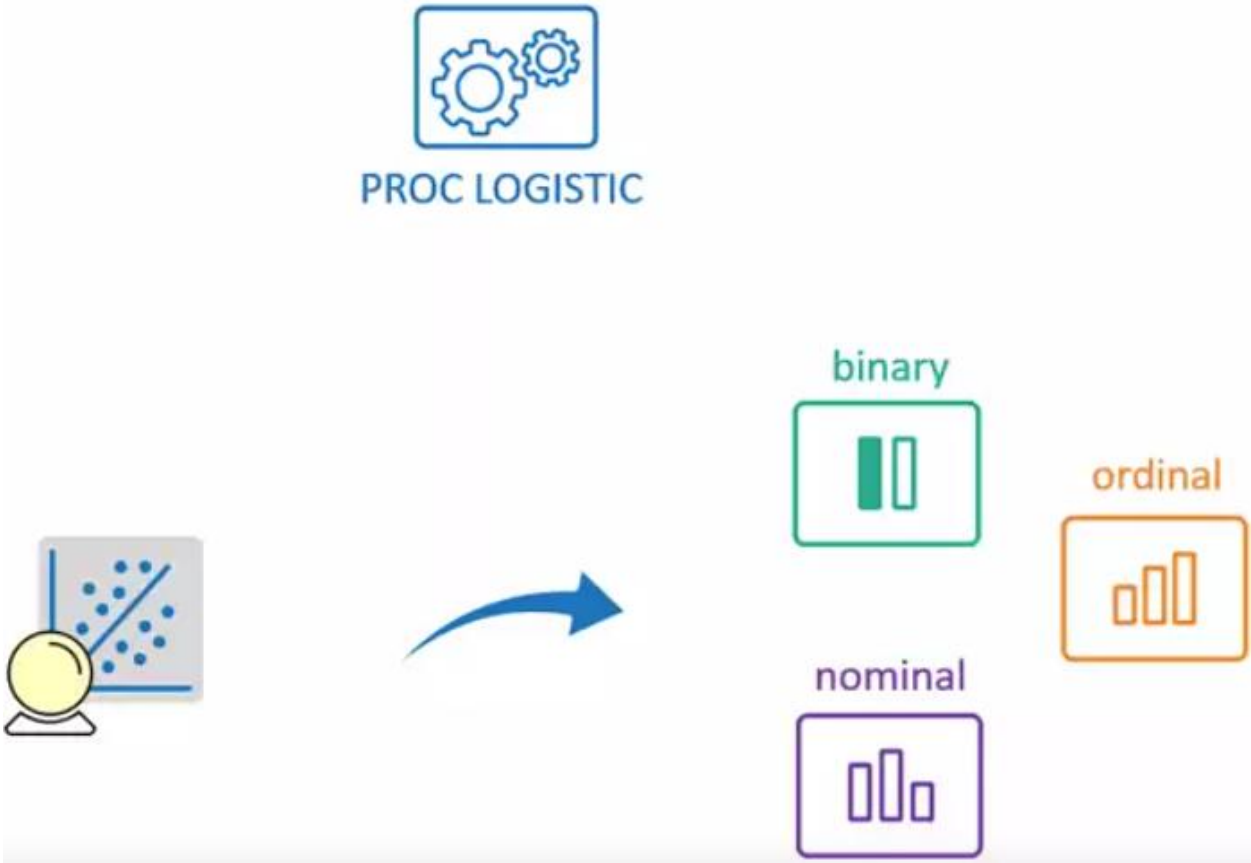

PROC LOGISTIC
maximum likelihood method
goodness-of-fit
concordant
discordant
tied



		Bought	
	$\hat{p}$	Higher	Lower
Did Not Buy	Higher	Tied	Discordant
	Lower	Concordant	Tied



Using PROC LOGISTIC to Fit Logistic Regression Models



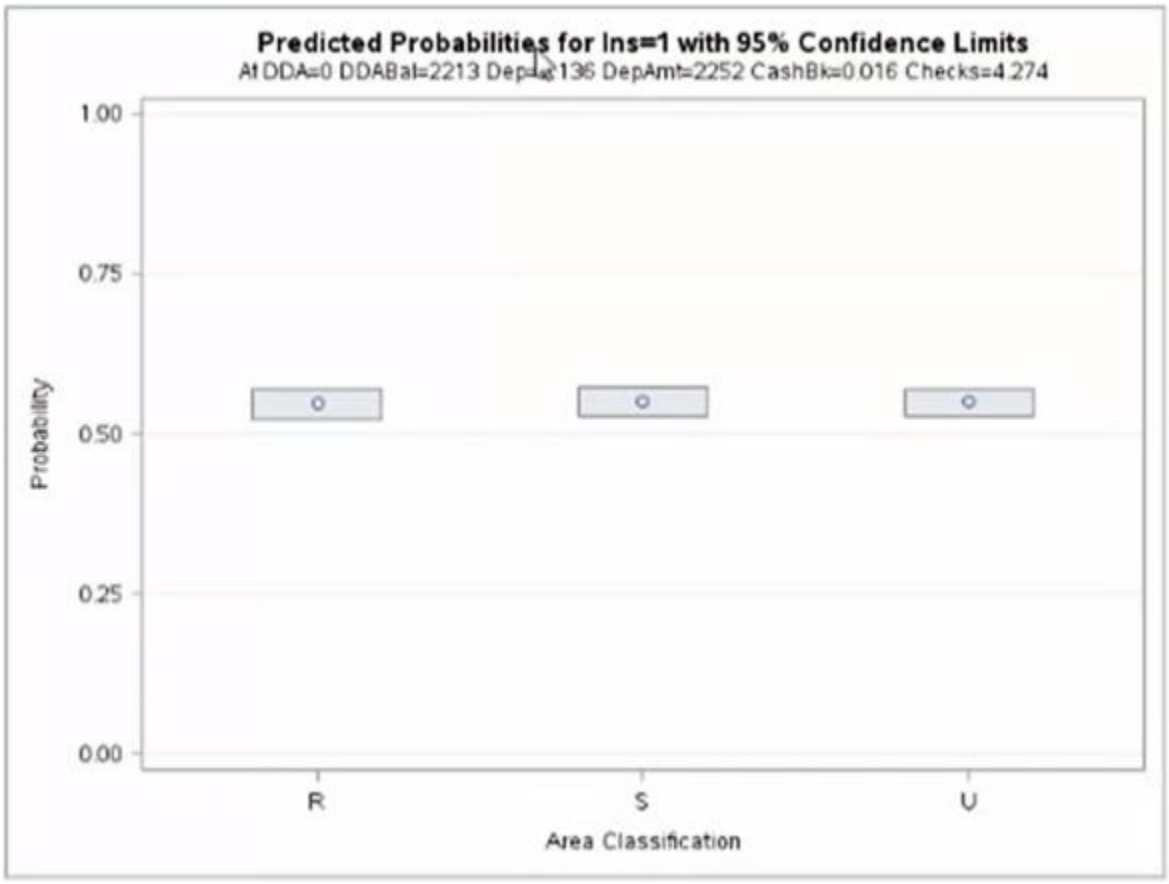
```
PROC LOGISTIC <options>;
    CLASS variable</v-options>;
    EFFECTPLOT <plot-type <(plot-definition-options)>>
                </options>;
    MODEL response=<effects></options>;
    ODDSRATIO <'label'> variable </options>;
    SCORE <options>;
    CODE <options>;
    UNITS <predictor1=list1> </option>;
RUN;
```

Demo Fitting a Basic Logistic Regression Model, Part 1

pmlr02d01.sas *

```
title1 "Logistic Regression Model for the Variable Annuity Data Set";
proc logistic data=work.train
              plots(only maxpoints=none)=
                   (effect(clband x=(ddabal depamt checks res))
                    oddsratio(type=horizontalstat));
   class res (param=ref ref='S') dda (param=ref ref='0');
   model ins(event='1')=dda ddabal dep depamt
                cashbk checks res / stb clodds=pl;
   units ddabal=1000 depamt=1000 / default=1;
   oddsratio 'Comparisons of Residential Classification' res /
             diff=all cl=pl;
   effectplot slicefit(sliceby=dda x=ddabal) / noobs;
   effectplot slicefit(sliceby=dda x=depamt) / noobs;
run;
title1 ;
```

```
title1 "Logistic Regression Model for the Variable Annuity Data Set";
proc logistic data=work.train
              plots(only maxpoints=none)=
                   (effect(clband x=(ddabal depamt checks res))
                    oddsratio (type=horizontalstat));
   class res (param=ref ref='S') dda (param=ref ref='0');
   model ins(event='1')=dda ddabal dep depamt
                checks res / stb clodds=pl;
   units ddabal=FIRST epamt=1000 / default=1;
   oddsratio 'Comparisons of Residential Classification' res /
             LAST cl=pl;
   effectplot slicefit(sliceby=dda x=ddabal) / noobs;
   effectplot slicefit(sliceby=dda x=depamt) / noobs;
run;
title1 ;
```



```
/* ================================================== */
/* Lesson 2, Section 1: l2d1.sas
   Demonstration: Fitting a Basic Logistic Regression Model,
   Parts 1 and 2
   [m642_1_k1, m642_1_k2; derived from pmlr02d01.sas]    */
/* ================================================== */

title1 "Logistic Regression Model for the Variable Annuity Data Set";
proc logistic data=work.train
          plots(only maxpoints=none)=(effect(clband x=(ddabal depamt checks res))
          oddsratio (type=horizontalstat));
   class res (param=ref ref='S') dda (param=ref ref='0');
   model ins(event='1')=dda ddabal dep depamt
           cashbk checks res / stb clodds=pl;
   units ddabal=1000 depamt=1000 / default=1;
   oddsratio 'Comparisons of Residential Classification' res / diff=all cl=pl;
   effectplot slicefit(sliceby=dda x=ddabal) / noobs;
   effectplot slicefit(sliceby=dda x=depamt) / noobs;
run;
title1;
```

Demo Fitting a Basic Logistic Regression Model, Part 2

Logistic Regression Model for the Variable Annuity Data Set

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TRAIN
Response Variable	Ins
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	21512
Number of Observations Used	21512

Response Profile		
Ordered Value	Ins	Total Frequency
1	0	14061
2	1	7451

Probability modeled is Ins=1.

Class Level Information			
Class	Value	Design Variables	
Res	R	1	0
	S	0	0
	U	0	1
DDA	0	0	
	1	1	

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	27759.675	26284.098
SC	27767.651	26355.885
-2 Log L	27757.675	26266.098

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1491.5772	8	<.0001
Score	1315.6105	8	<.0001
Wald	1256.8282	8	<.0001

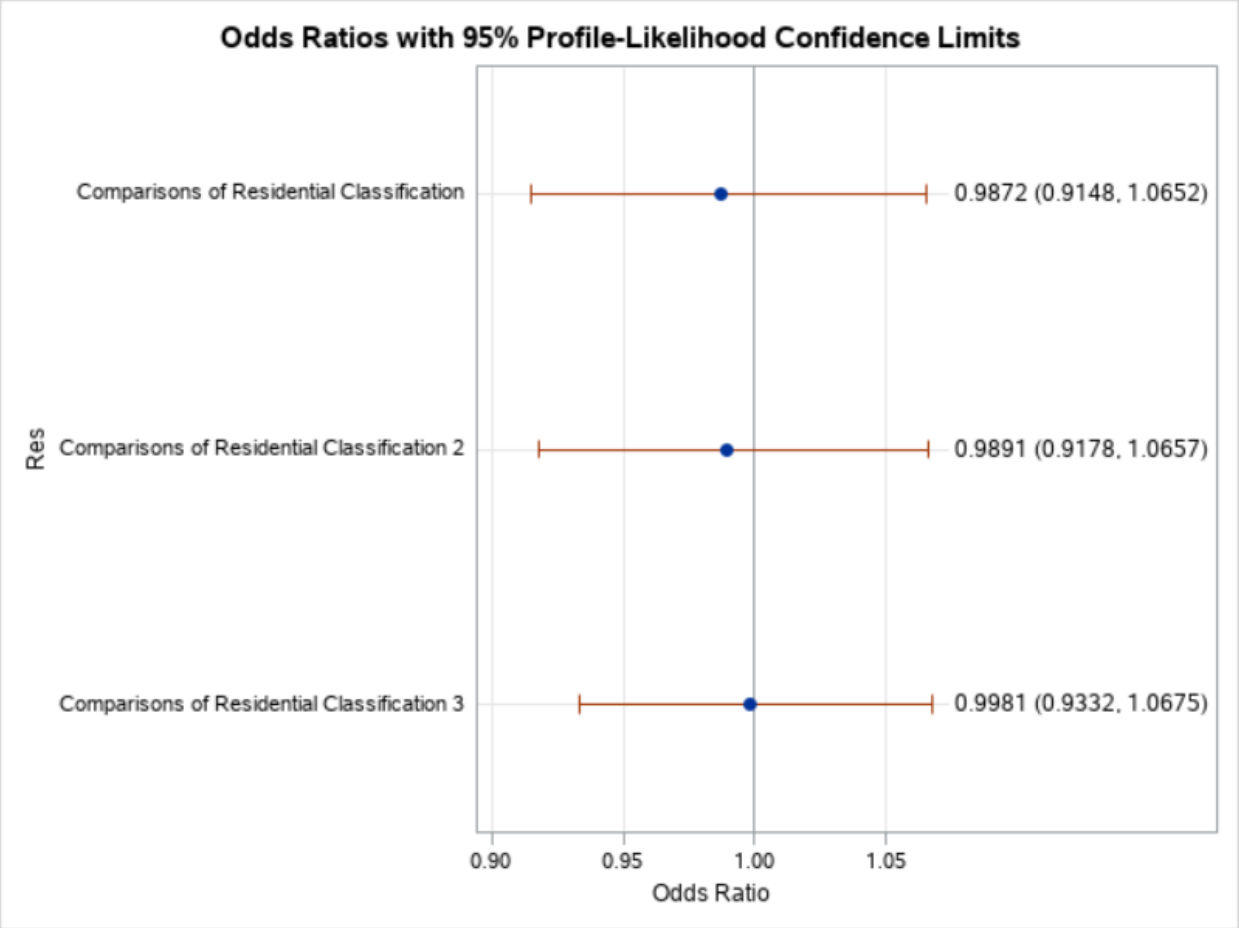
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1491.5772	8	<.0001
Score	1315.6105	8	<.0001
Wald	1256.8282	8	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
DDA	1	484.0020	<.0001
DDABal	1	317.1284	<.0001
Dep	1	26.0277	<.0001
DepAmt	1	10.1271	0.0015
CashBk	1	19.8706	<.0001
Checks	1	0.0309	0.8604
Res	2	0.1229	0.9404

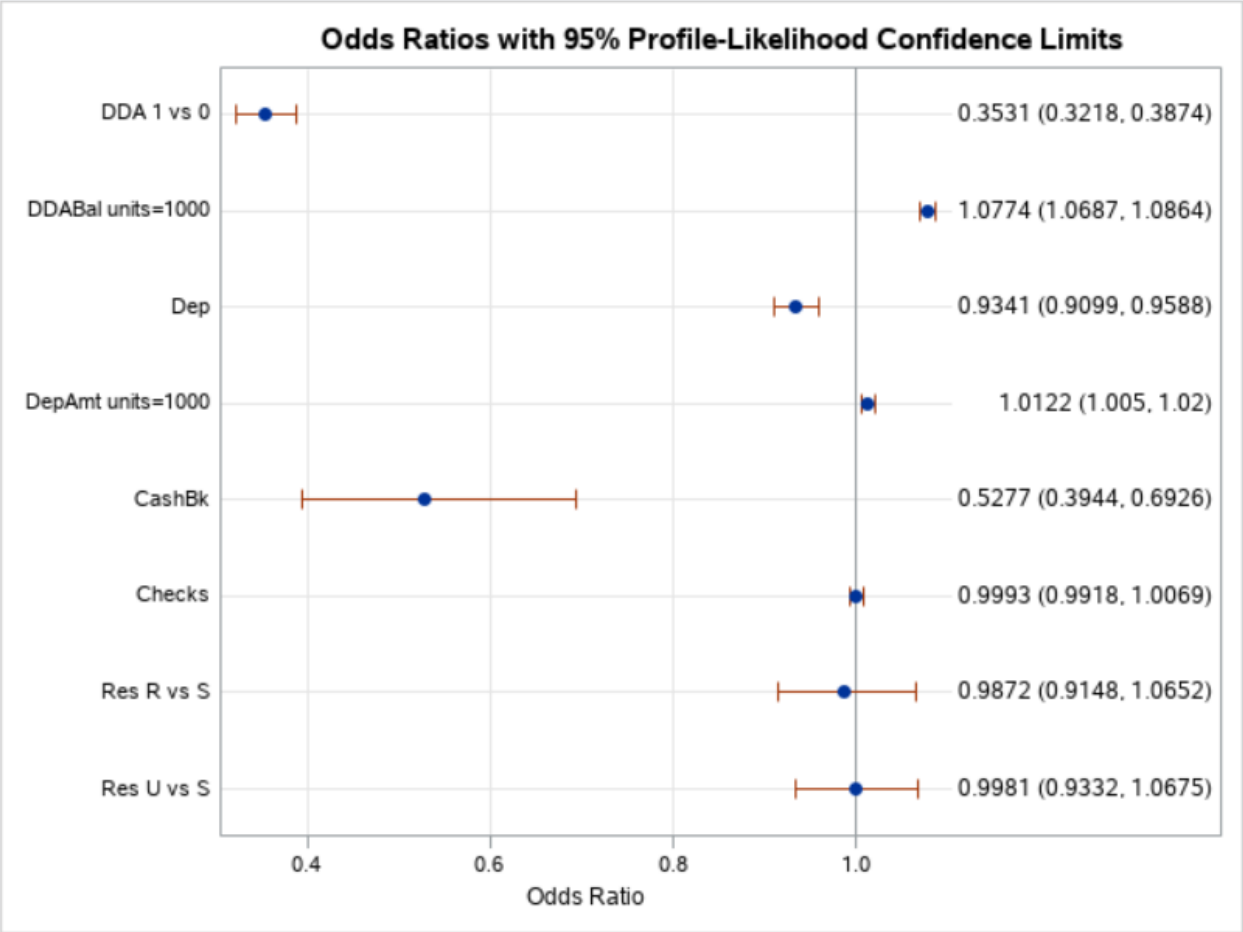
Analysis of Maximum Likelihood Estimates							
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept		1	0.1706	0.0374	20.8591	<.0001	
DDA	1	1	-1.0410	0.0473	484.0020	<.0001	-0.2226
DDABal		1	0.000075	4.188E-6	317.1284	<.0001	0.3135
Dep		1	-0.0682	0.0134	26.0277	<.0001	-0.0648
DepAmt		1	0.000012	3.819E-6	10.1271	0.0015	0.0460
CashBk		1	-0.6393	0.1434	19.8706	<.0001	-0.0468
Checks		1	-0.00068	0.00384	0.0309	0.8604	-0.00193
Res	R	1	-0.0129	0.0388	0.1106	0.7395	-0.00308
Res	U	1	-0.00191	0.0343	0.0031	0.9557	-0.00051

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	67.2	Somers' D	0.357
Percent Discordant	31.5	Gamma	0.362
Percent Tied	1.3	Tau-a	0.162
Pairs	104768511	c	0.679

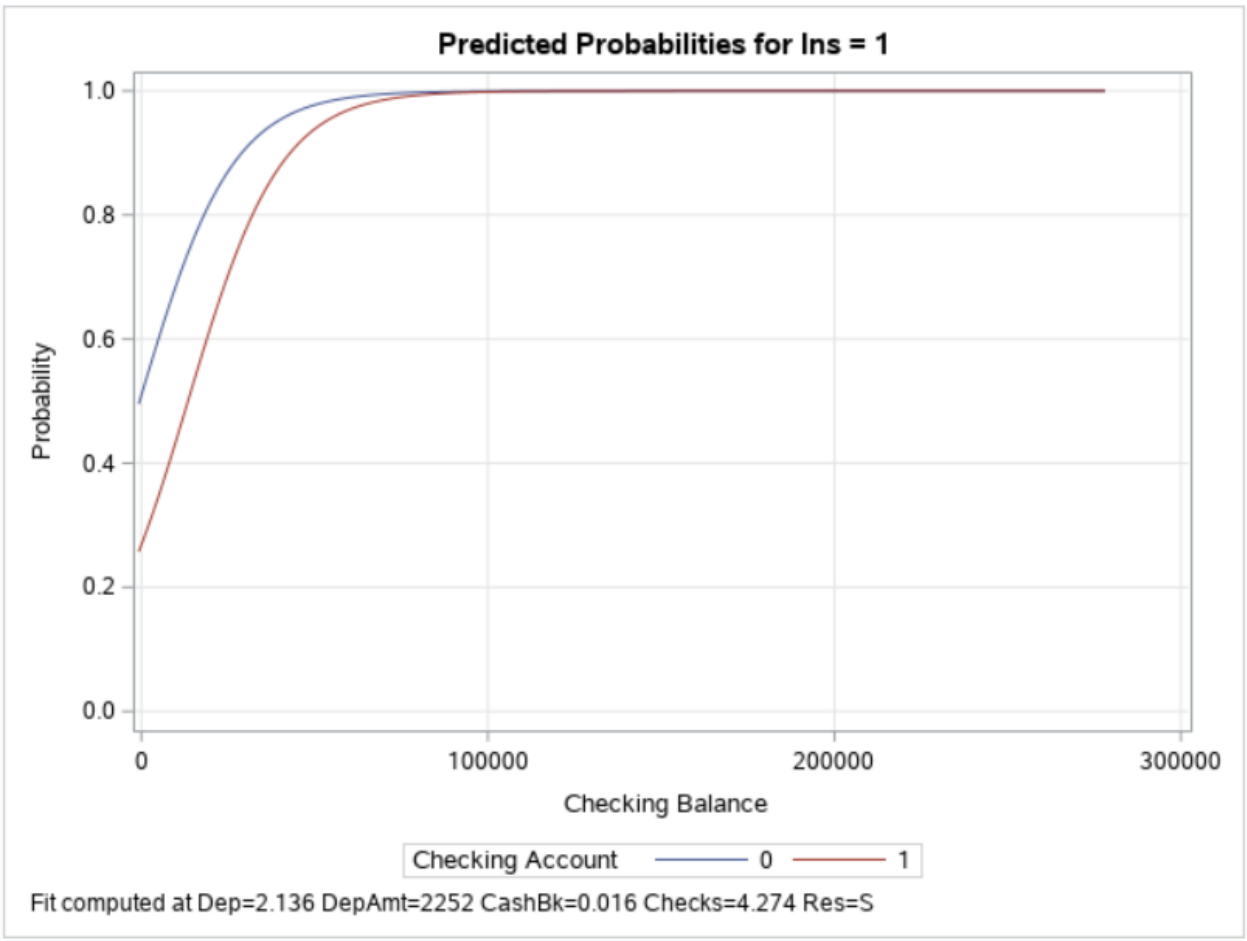
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Label	Odds Ratio	Estimate	95% Confidence Limits	
Comparisons of Residential Classification	Res R vs S	0.987	0.915	1.065
Comparisons of Residential Classification 2	Res R vs U	0.989	0.918	1.066
Comparisons of Residential Classification 3	Res U vs S	0.998	0.933	1.068

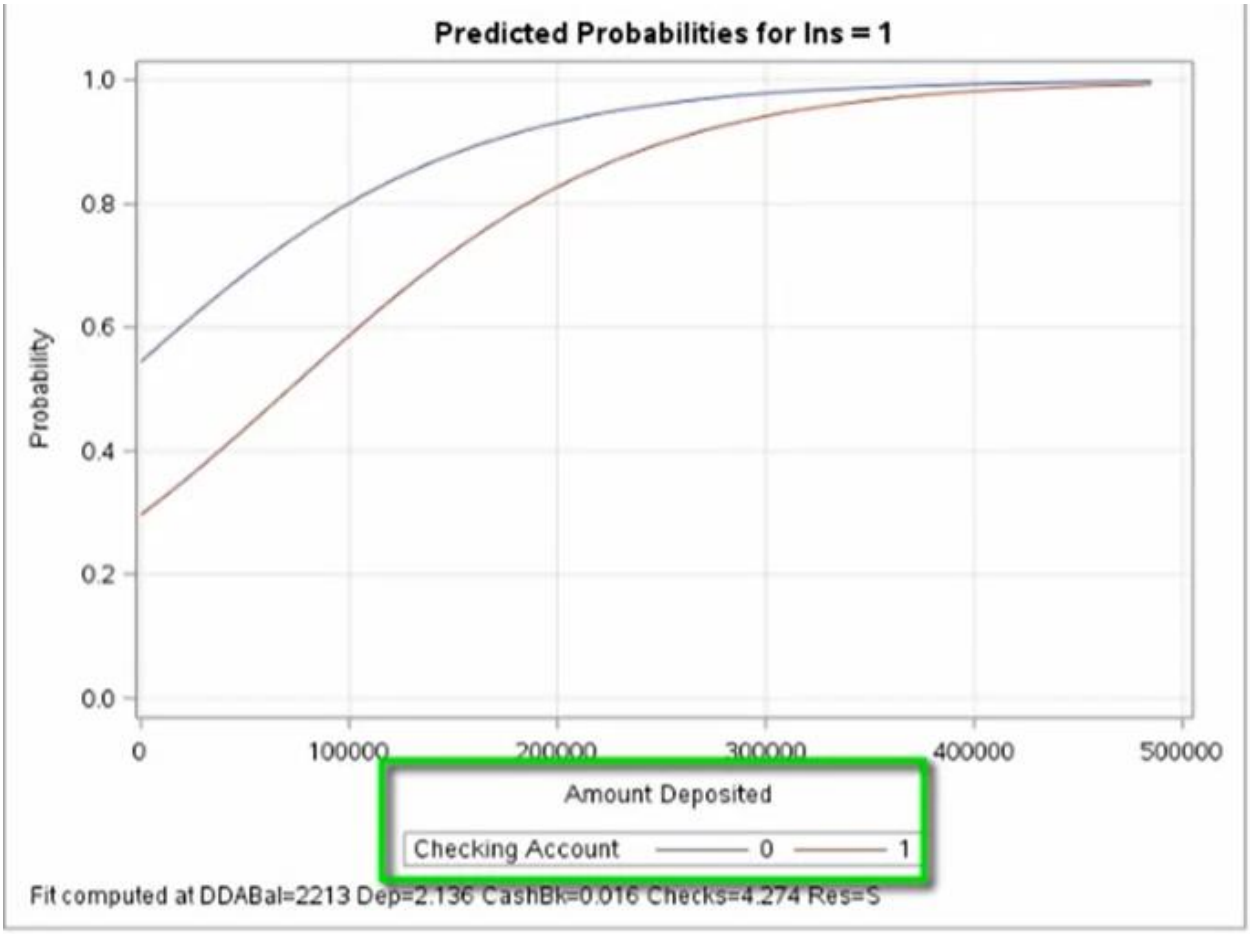


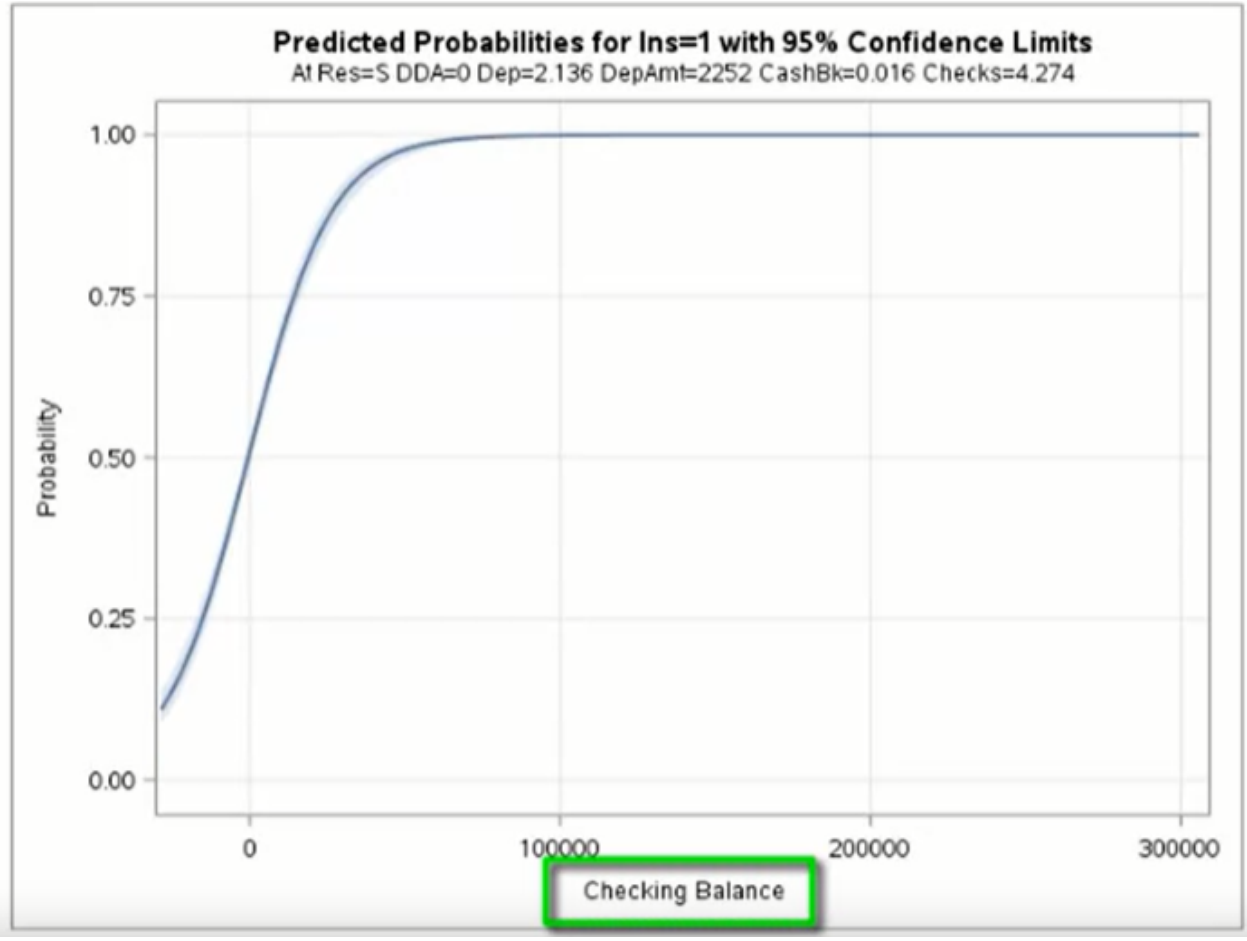
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
DDA 1 vs 0	1.0000	0.353	0.322	0.387
DDABal	1000.0	1.077	1.069	1.086
Dep	1.0000	0.934	0.910	0.959
DepAmt	1000.0	1.012	1.005	1.020
CashBk	1.0000	0.528	0.394	0.693
Checks	1.0000	0.999	0.992	1.007
Res R vs S	1.0000	0.987	0.915	1.065
Res U vs S	1.0000	0.998	0.933	1.068

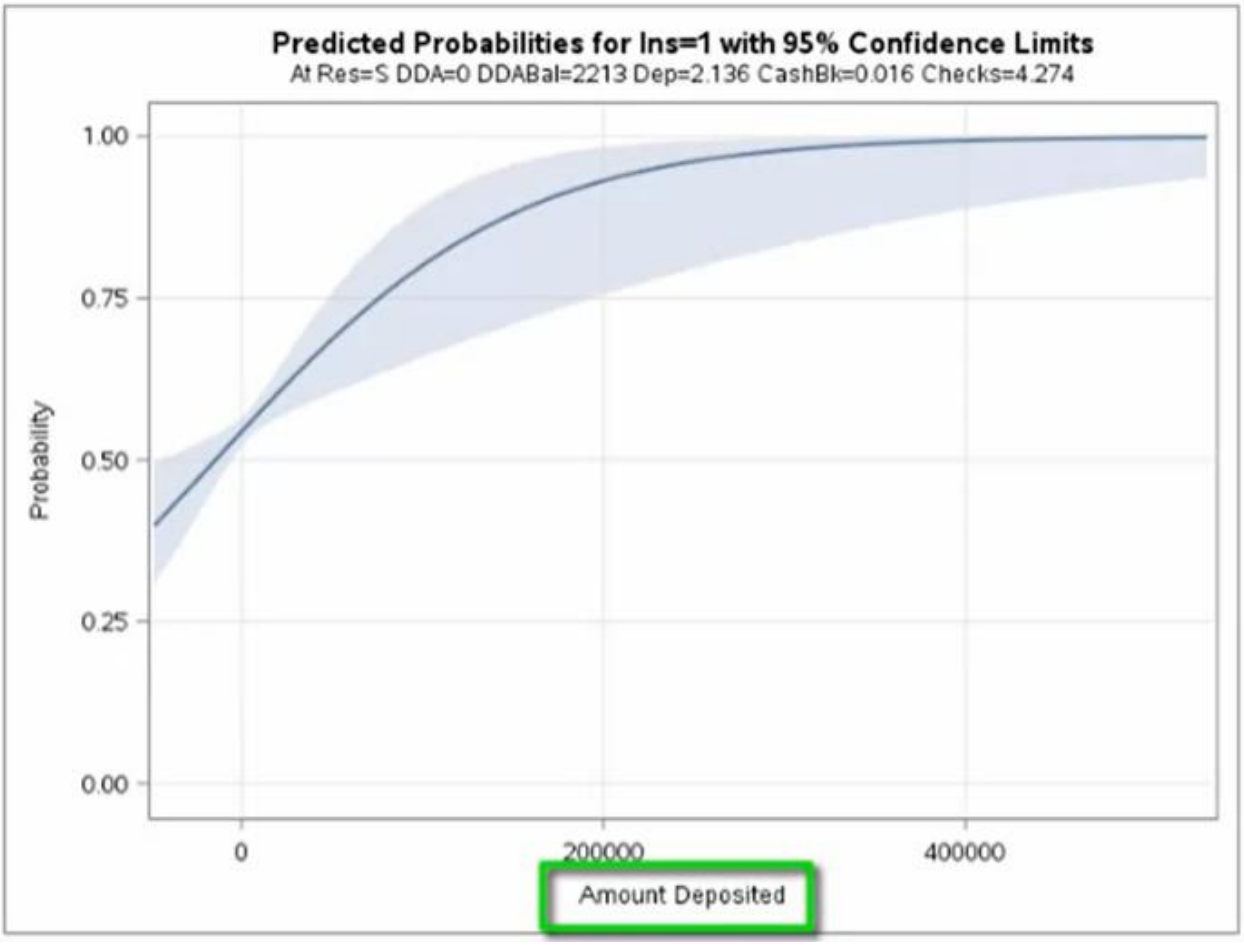


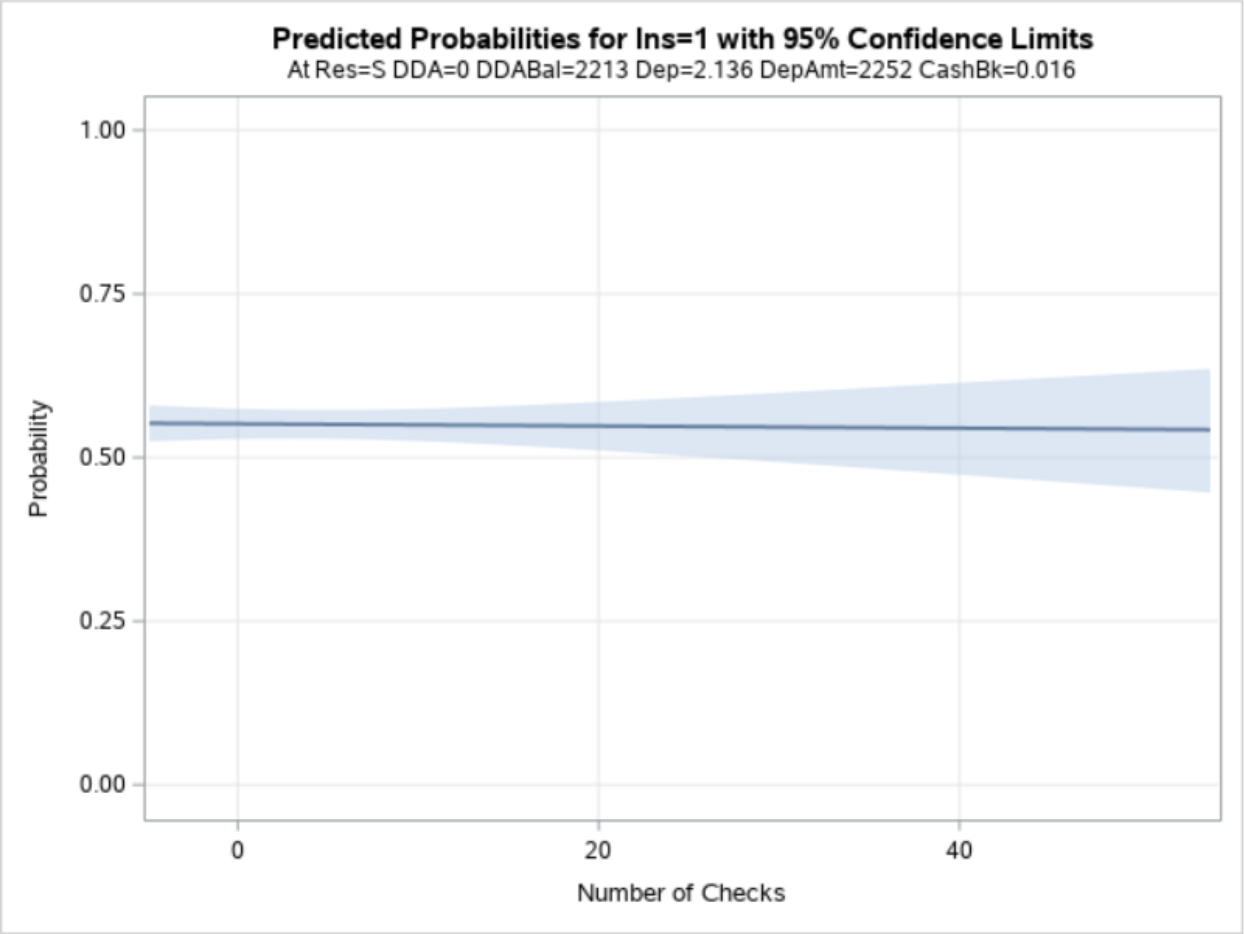
The Effects that do not cross value 1 are significant variables such as DDA 1 vs 0, DDABal, Dep, DepAmt, and CashBk.

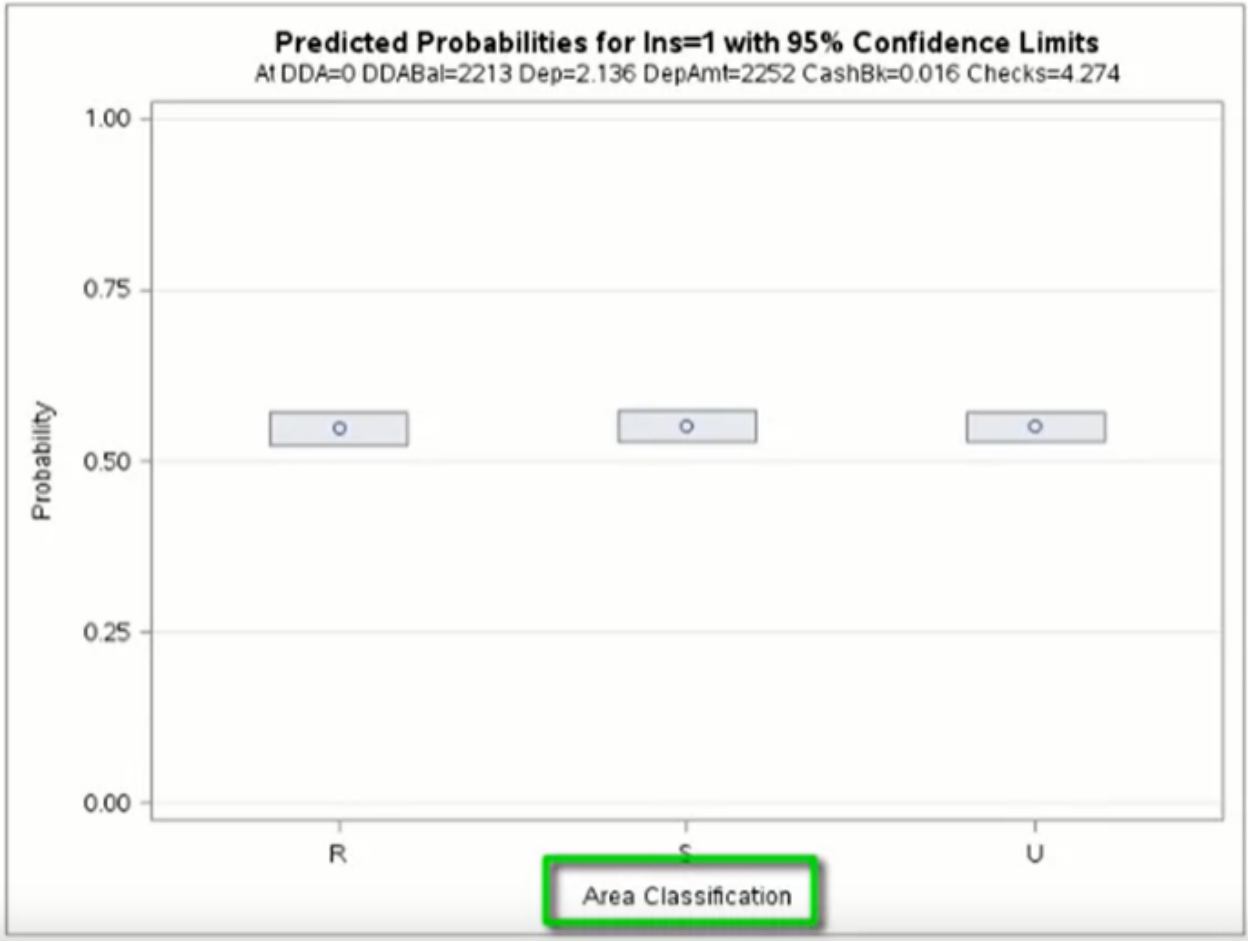

Predicted Probabilities for Ins = 1
1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
100000
200000
300000
Checking Balance
Checking Account
0
1
Fit computed at Dep=2.136 DepAmt=2252 CashBk=0.016 Checks=4.274 Res=S


Predicted Probabilities for Ins = 1
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
100000
200000
300000
400000
500000
Amount Deposited
Checking Account
0
1
Fit computed at DDABal=2213 Dep=2.136 CashBk=0.016 Checks=4.274 Res=S


Predicted Probabilities for Ins=1 with 95% Confidence Limits
At Res=S DDA=0 Dep=2.136 DepAmt=2252 CashBk=0.016 Checks=4.274
1.00
0.75
0.50
0.25
0.00
Probability
0
100000
200000
300000
Checking Balance


Predicted Probabilities for Ins=1 with 95% Confidence Limits
At Res=S DDA=0 DDABal=2213 Dep=2.136 CashBk=0.016 Checks=4.274
1.00
0.75
0.50
0.25
0.00
Probability
0
200000
400000
Amount Deposited


Predicted Probabilities for Ins=1 with 95% Confidence Limits
At Res=S DDA=0 DDABal=2213 Dep=2.136 DepAmt=2252 CashBk=0.016
1.00
0.75
0.50
0.25
0.00
Probability
0
20
40
Number of Checks



Question 2.02

Given the following Analysis of Maximum Likelihood Estimates table, what are the first, second, and third most powerful predictors based on the standardized estimates?

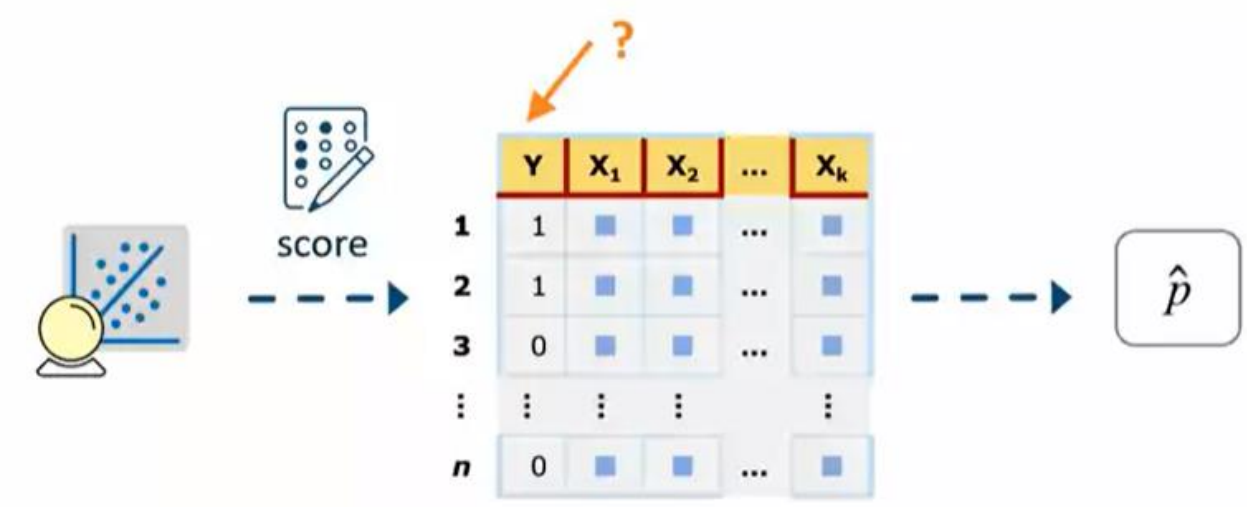
Analysis of Maximum Likelihood Estimates							
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**	**Standardized Estimate**
Intercept		1	0.1519	0.0304	24.8969	<.0001	
DDA		1	-0.9699	0.0385	633.6092	<.0001	-0.2074
DDABal		1	0.000072	3.39E-6	448.5968	<.0001	0.2883
Dep		1	-0.0714	0.0109	42.8222	<.0001	-0.0678
DepAmt		1	0.000018	3.225E-6	30.6227	<.0001	0.0660
CashBk		1	-0.5629	0.1145	24.1615	<.0001	-0.0408
Checks		1	-0.00402	0.00317	1.6168	0.2035	-0.0114
Res	**R**	1	-0.0467	0.0316	2.1907	0.1388	
Res	**U**	1	-0.0379	0.0280	1.8375	0.1752	

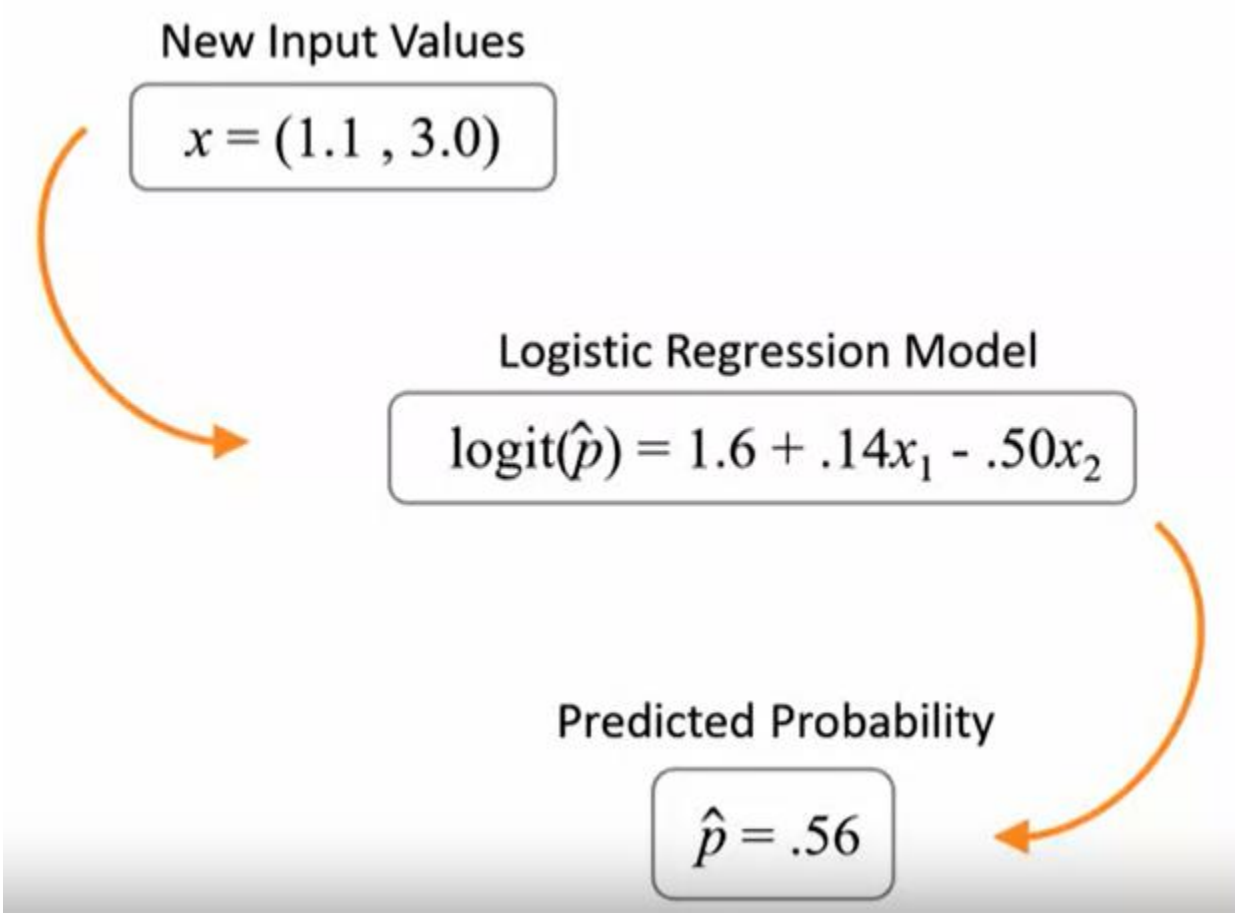
DDA, **DDABal**, and **Dep**

Correct

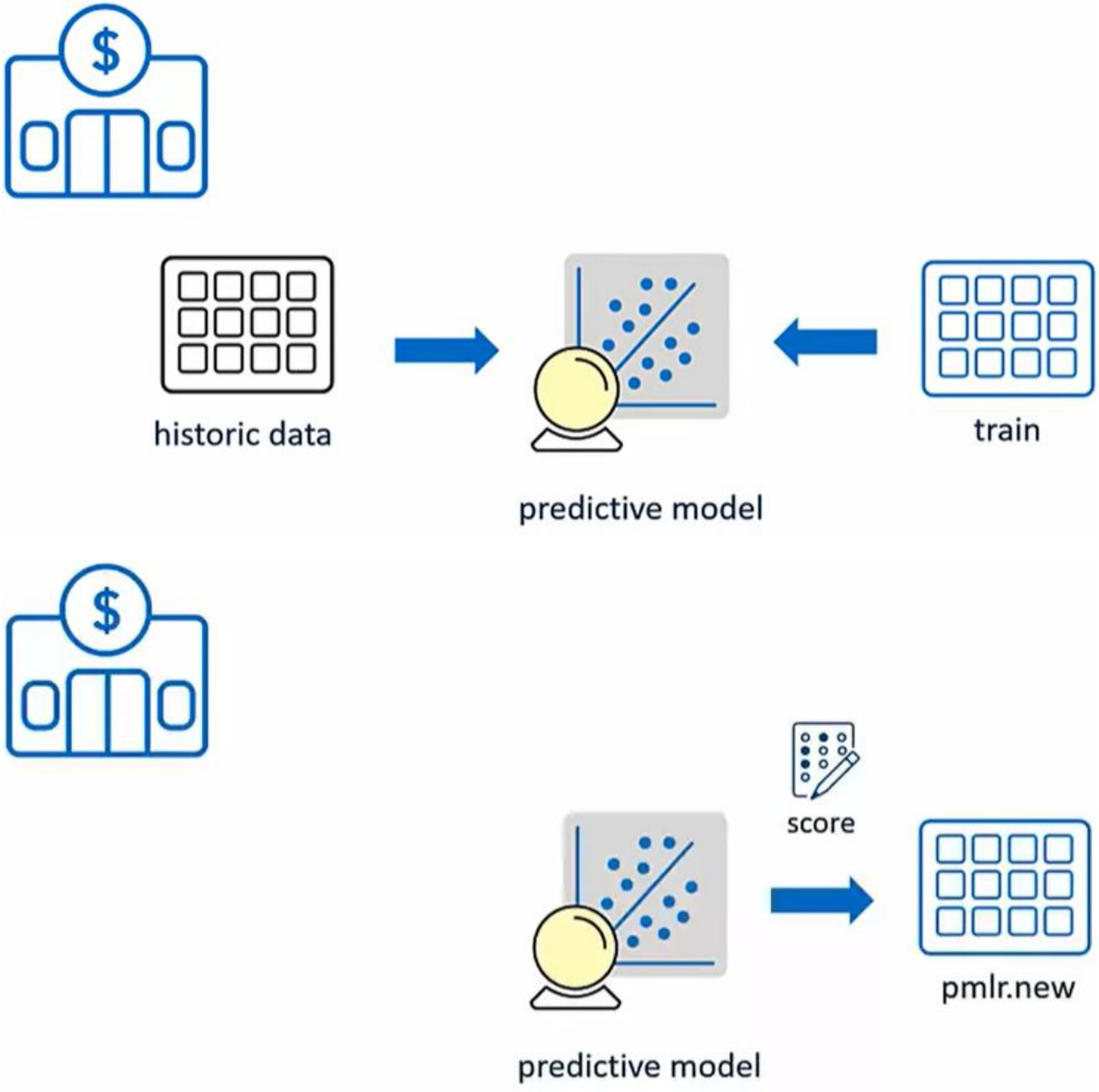
The answer is based on the rank order of the absolute values of the standardized estimates.

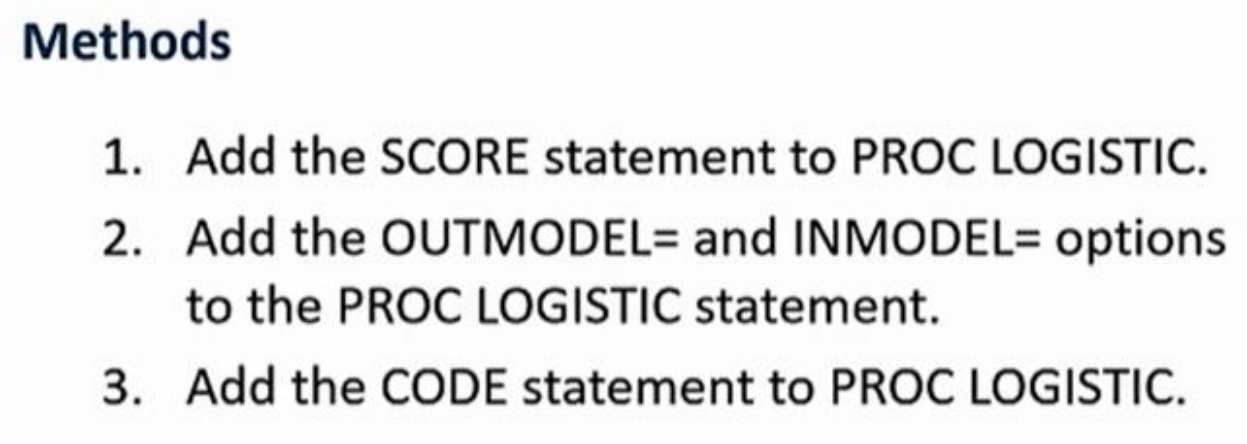
Scoring New Cases





Demo Scoring New Cases






```
/* Run this code before demo l2d2 */

/* ================================================== */
/* Lesson 1, Section 1: l1d1.sas
   Demonstration: Examining the Code for Generating
   Descriptive Statistics and Frequency Tables        */
/* ================================================== */

data work.develop;
  set pmlr.develop;
run;

%global inputs;
%let inputs=ACCTAGE DDA DDABAL DEP DEPAMT CASHBK
        CHECKS DIRDEP NSF NSFAMT PHONE TELLER
        SAV SAVBAL ATM ATMAMT POS POSAMT CD
        CDBAL IRA IRABAL LOC LOCBAL INV
        INVBAL ILS ILSBAL MM MMBAL MMCRED MTG
        MTGBAL CC CCBAL CCPURC SDB INCOME
        HMOWN LORES HMVAL AGE CRSCORE MOVED
        INAREA;
```

```
proc means data=work.develop n nmiss mean min max;
  var &inputs;
run;

proc freq data=work.develop;
  tables ins branch res;
run;

/* ---------------------------------------------------- */
/* Lesson 1, Section 2: l1d2.sas
  Demonstration: Splitting the Data                  */
/* ---------------------------------------------------- */


/* Sort the data by the target in preparation for stratified sampling. */

proc sort data=work.develop out=work.develop_sort;
  by ins;
run;

/* The SURVEYSELECT procedure will perform stratified sampling
  on any variable in the STRATA statement. The OUTALL option
  specifies that you want a flag appended to the file to
  indicate selected records, not simply a file comprised
  of the selected records. */

proc surveyselect noprint data=work.develop_sort
            samprate=.6667 stratumseed=restore
```

```
            out=work.develop_sample
            seed=44444 outall;
  strata ins;
run;

/* Verify stratification. */

proc freq data=work.develop_sample;
  tables ins*selected;
run;

/* Create training and validation data sets. */

data work.train(drop=selected SelectionProb SamplingWeight)
    work.valid(drop=selected SelectionProb SamplingWeight);
  set work.develop_sample;
  if selected then output work.train;
  else output work.valid;
run;




/* ================================================== */
/* Lesson 2, Section 1: l2d1.sas
  Demonstration: Fitting a Basic Logistic
  Regression Model, Parts 1 and 2                */
/* ================================================== */

title1 "Logistic Regression Model for the Variable Annuity Data Set";
```

```
proc logistic data=work.train
          plots(only maxpoints=none)=(effect(clband x=(ddabal depamt checks res))
          oddsratio (type=horizontalstat));
  class res (param=ref ref='S') dda (param=ref ref='0');
  model ins(event='1')=dda ddabal dep depamt
           cashbk checks res / stb clodds=pl;
  units ddabal=1000 depamt=1000 / default=1;
  oddsratio 'Comparisons of Residential Classification' res / diff=all cl=pl;
  effectplot slicefit(sliceby=dda x=ddabal) / noobs;
  effectplot slicefit(sliceby=dda x=depamt) / noobs;
run;
title1;

/* ================================================== */
/* Lesson 2, Section 1: l2d2.sas
  Demonstration: Scoring New Cases
  [m642_1_n; derived from pmlr02d02.sas]            */
/* ================================================== */

/* Score a new data set with one run of the LOGISTIC procedure with the
  SCORE statement.  */

proc logistic data=work.train noprint;
  class res (param=ref ref='S');
  model ins(event='1')= res dda ddabal dep depamt cashbk checks;
  score data = pmlr.new out=work.scored1;
run;

title1 "Predicted Probabilities from Scored Data Set";
```

```
proc print data=work.scored1(obs=10);
  var p_1 dda ddabal dep depamt cashbk checks res;
run;
```

Predicted Probabilities from Scored Data Set

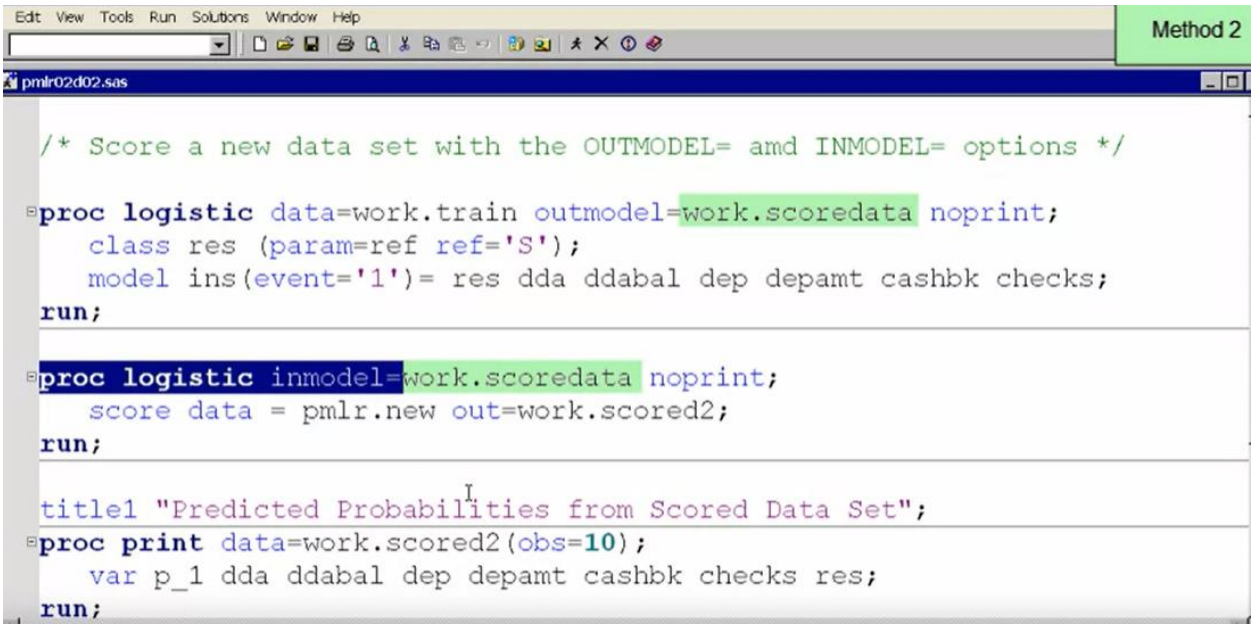
Obs	P_1	DDA	DDABal	Dep	DepAmt	CashBk	Checks	Res
1	0.27023	1	56.29	2	955.51	0	1	U
2	0.32036	1	3292.17	2	961.60	0	1	U
3	0.29822	1	1723.86	2	2108.65	0	2	U
4	0.54208	0	0.00	0	0.00	0	0	U
5	0.26946	1	67.91	2	519.24	0	3	S
6	0.32228	1	2554.58	1	501.36	0	2	S
7	0.27038	1	0.00	2	2883.08	0	12	R
8	0.30399	1	2641.33	3	4521.61	0	8	S
9	0.54255	0	0.00	0	0.00	0	0	S
10	0.27956	1	52.22	1	75.59	0	0	R

```
title1 "Mean of Predicted Probabilities from Scored Data Set";
proc means data=work.scored1 mean nolabels;
  var p_1;
run;
```

Mean of Predicted Probabilities from Scored Data Set

The MEANS Procedure

Analysis Variable : P_1
Mean
0.3448203




```
/* Score a new data set with the OUTMODEL= amd INMODEL= options */

proc logistic data=work.train outmodel=work.scoredata noprint;
  class res (param=ref ref='S');
  model ins(event='1')= res dda ddabal dep depamt cashbk checks;
run;

proc logistic inmodel=work.scoredata noprint;
  score data = pmlr.new out=work.scored2;
run;

title1 "Predicted Probabilities from Scored Data Set";
proc print data=work.scored2(obs=10);
  var p_1 dda ddabal dep depamt cashbk checks res;
run;
```

Predicted Probabilities from Scored Data Set

Obs	P_1	DDA	DDABal	Dep	DepAmt	CashBk	Checks	Res
1	0.27023	1	56.29	2	955.51	0	1	U
2	0.32036	1	3292.17	2	961.60	0	1	U
3	0.29822	1	1723.86	2	2108.65	0	2	U
4	0.54208	0	0.00	0	0.00	0	0	U
5	0.26946	1	67.91	2	519.24	0	3	S
6	0.32228	1	2554.58	1	501.36	0	2	S
7	0.27038	1	0.00	2	2883.08	0	12	R
8	0.30399	1	2641.33	3	4521.61	0	8	S
9	0.54255	0	0.00	0	0.00	0	0	S
10	0.27956	1	52.22	1	75.59	0	0	R

Method 3

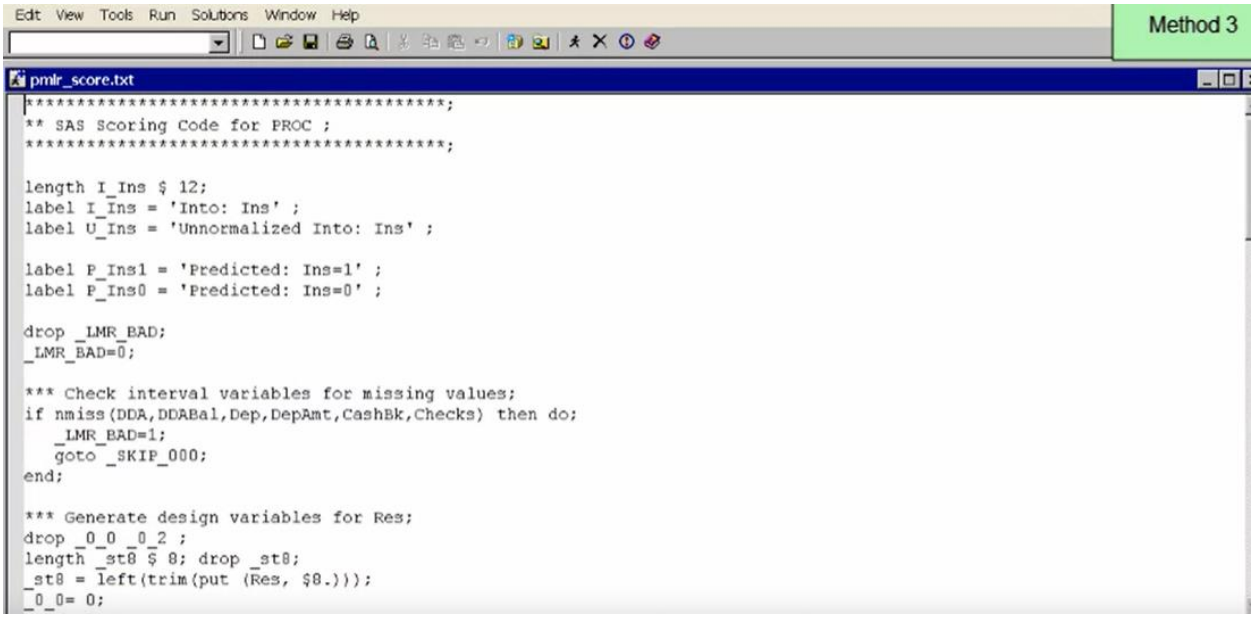
pmlr02d02.sas

```
/* Score a new data set with the CODE Statement  */

proc logistic data=work.train noprint;
   class res (param=ref ref='S');
   model ins(event='1')= res dda ddabal dep depamt cashbk checks;
   code file="&PMLRfolder\pmlr_score.txt";
run;

data work.scored3;
   set pmlr.new;
   %include "&PMLRfolder\pmlr_score.txt";
run;

title1 "Predicted Probabilities from Scored Data Set";
proc print data=work.scored3(obs=10);
   var p_ins1 dda ddabal dep depamt cashbk checks res;
```


Edit View Tools Run Solutions Window Help
Method 3
pmlr_score.txt
**************************************;
** SAS Scoring Code for PROC ;
**************************************;
length I_Ins $ 12;
label I_Ins = 'Into: Ins' ;
label U_Ins = 'Unnormalized Into: Ins' ;
label P_Ins1 = 'Predicted: Ins=1' ;
label P_Ins0 = 'Predicted: Ins=0' ;
drop _LMR_BAD;
_LMR_BAD=0;
*** Check interval variables for missing values;
if nmiss(DDA,DDABal,Dep,DepAmt,CashBk,Checks) then do;
_LMR_BAD=1;
goto _SKIP_000;
end;
*** Generate design variables for Res;
drop _0_0 _0_2 ;
length _st8 $ 8; drop _st8;
_st8 = left(trim(put (Res, $8.)));
_0_0= 0;

pmlr_score.txt

```
*** Effect: DepAmt;
_LP0 = _LP0 + (0.0000121517215) * DepAmt;
*** Effect: CashBk;
_LP0 = _LP0 + (-0.63930659872598) * CashBk;
*** Effect: Checks;
_LP0 = _LP0 + (-0.00067554471582) * Checks;

*** Predicted values;
drop _MAXP _IY _P0 _P1;
_TEMP = 0.17060673004991  + _LP0;
if (_TEMP < 0) then do;
   _TEMP = exp(_TEMP);
   _P0 = _TEMP / (1 + _TEMP);
end;
else _P0 = 1 / (1 + exp(-_TEMP));
_P1 = 1.0 - _P0;
P_Ins1 = _P0;
_MAXP = _P0;
_IY = 1;
P_Ins0 = _P1;
if (_P1 >  _MAXP + 1E-8) then do;
   _MAXP = _P1;
   _IY = 2;
end;
select( _IY );
```

data work.scored3;

set pmlr.new;

%include "&PMLRfolder/pmlr_score.txt";

run;

title1 "Predicted Probabilities from Scored Data Set";

proc print data=work.scored3(obs=10);

var p_ins1 dda ddabal dep depamt cashbk checks res;

run;

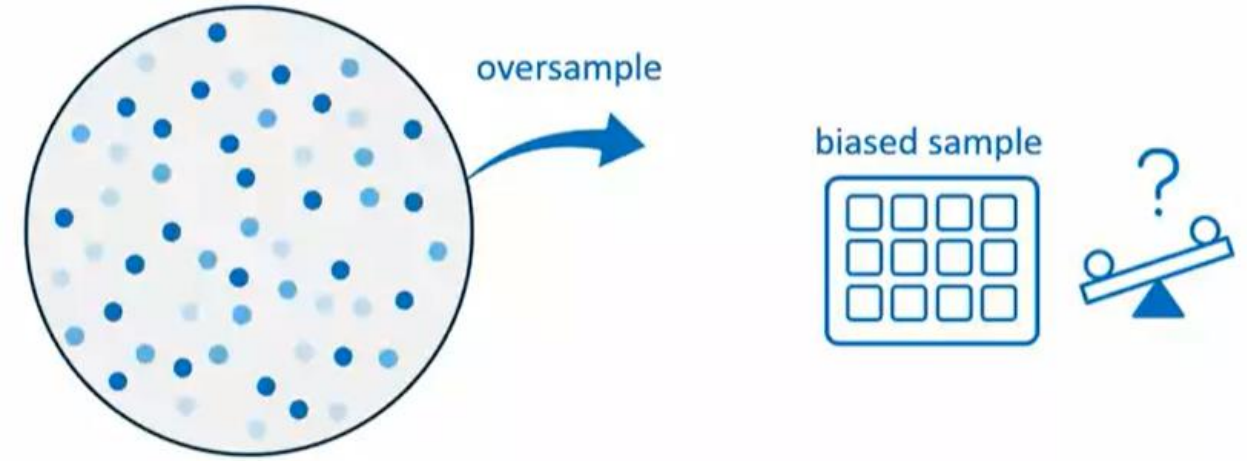
title1 ;

Predicted Probabilities from Scored Data Set

Obs	P_Ins1	DDA	DDABal	Dep	DepAmt	CashBk	Checks	Res
1	0.27023	1	56.29	2	955.51	0	1	U
2	0.32036	1	3292.17	2	961.60	0	1	U
3	0.29822	1	1723.86	2	2108.65	0	2	U
4	0.54208	0	0.00	0	0.00	0	0	U
5	0.26946	1	67.91	2	519.24	0	3	S
6	0.32228	1	2554.58	1	501.36	0	2	S
7	0.27038	1	0.00	2	2883.08	0	12	R
8	0.30399	1	2641.33	3	4521.61	0	8	S
9	0.54255	0	0.00	0	0.00	0	0	S
10	0.27956	1	52.22	1	75.59	0	0	R

Correcting for Oversampling

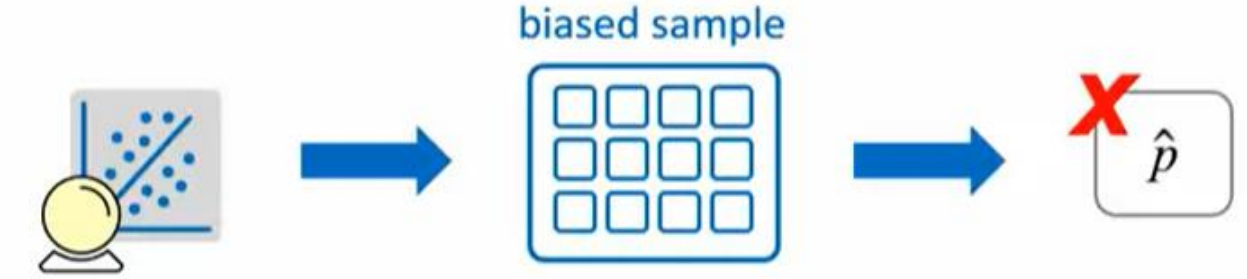
Introduction

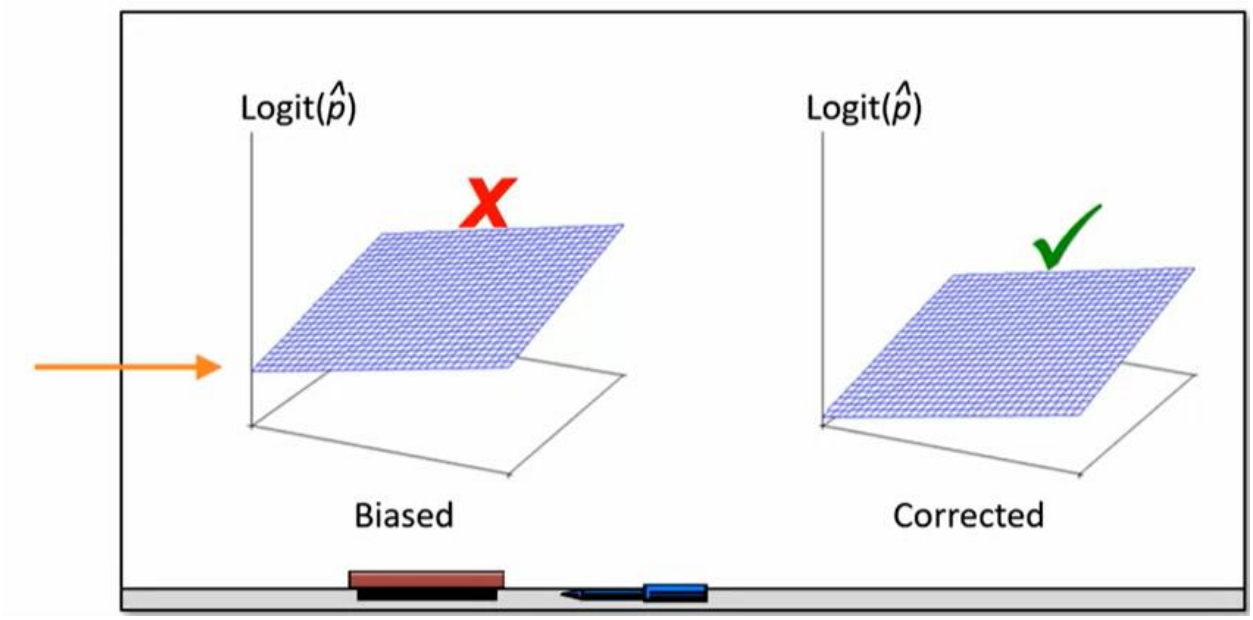


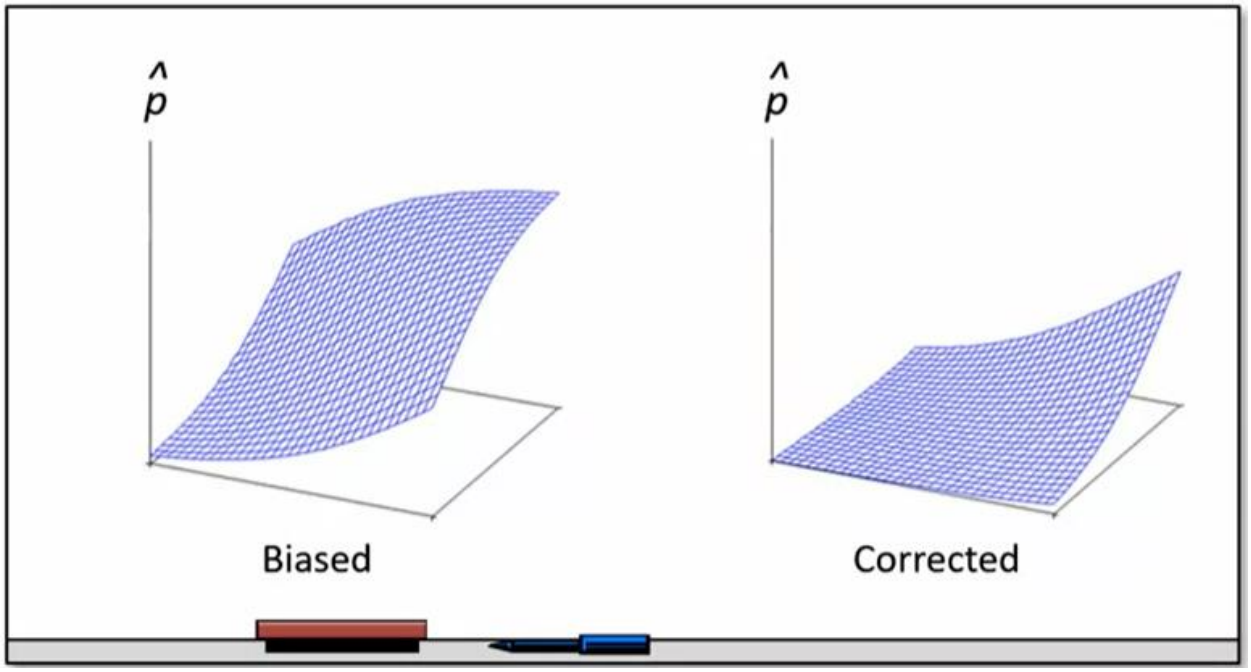
In this topic, you learn to do the following:

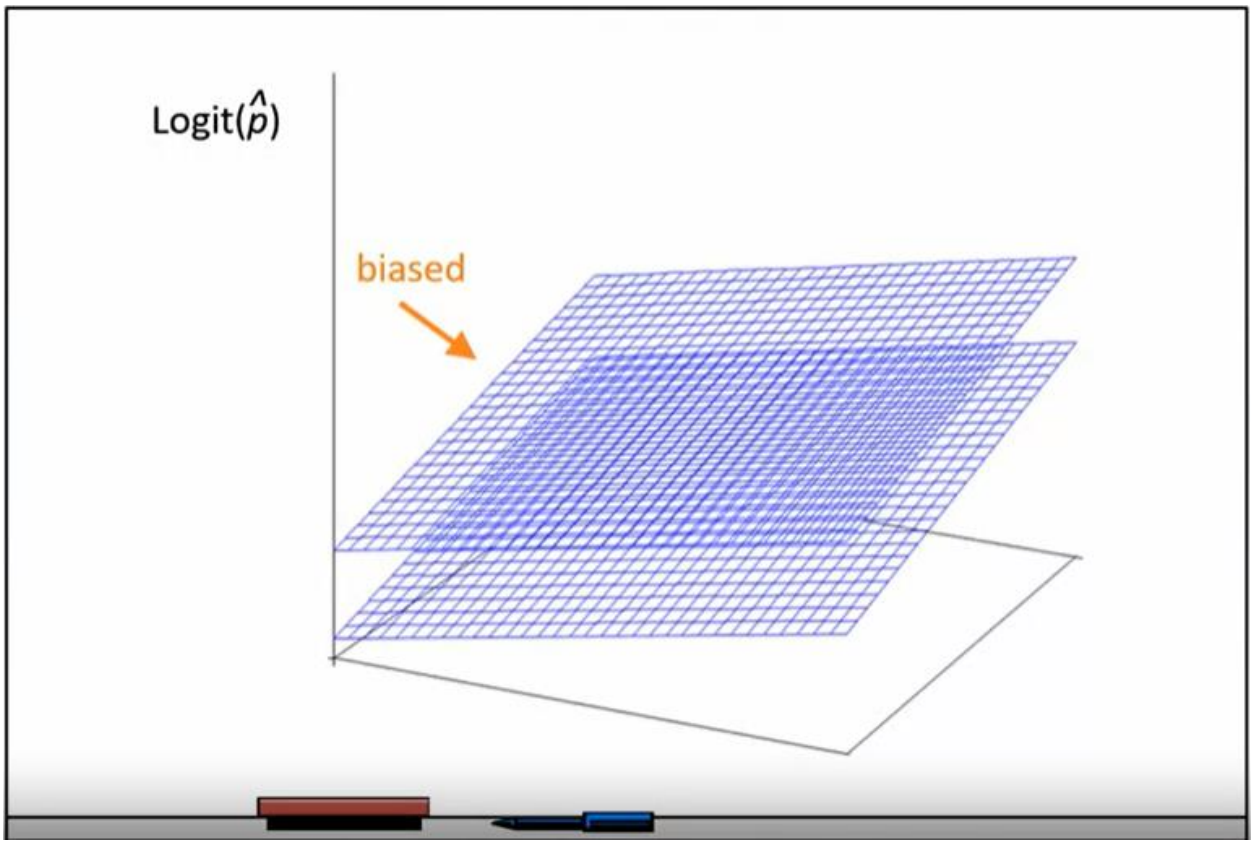
- describe the effect of oversampling
- describe the offset method of adjusting for oversampling
- adjust a model for oversampling in PROC LOGISTIC

Understanding the Effect of Oversampling

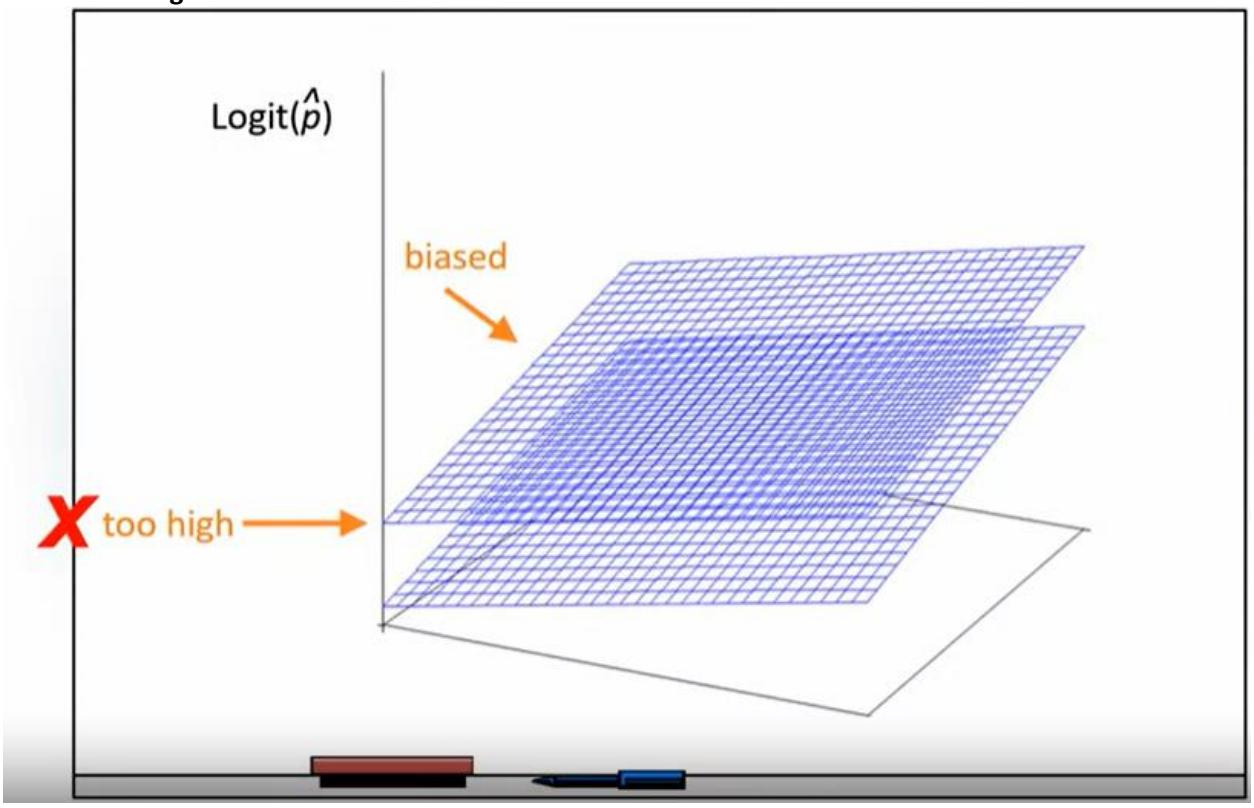


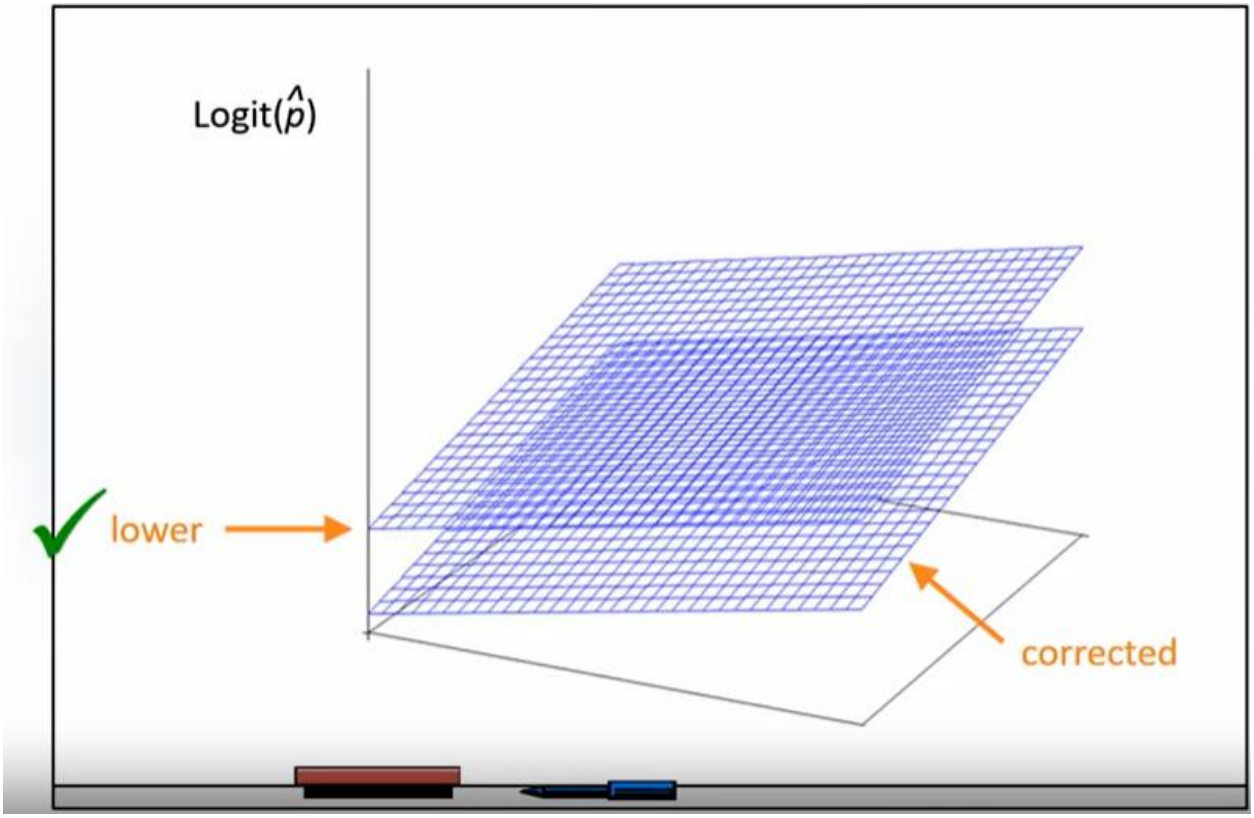

Logit($\hat{p}$)
Logit($\hat{p}$)
Biased
Corrected

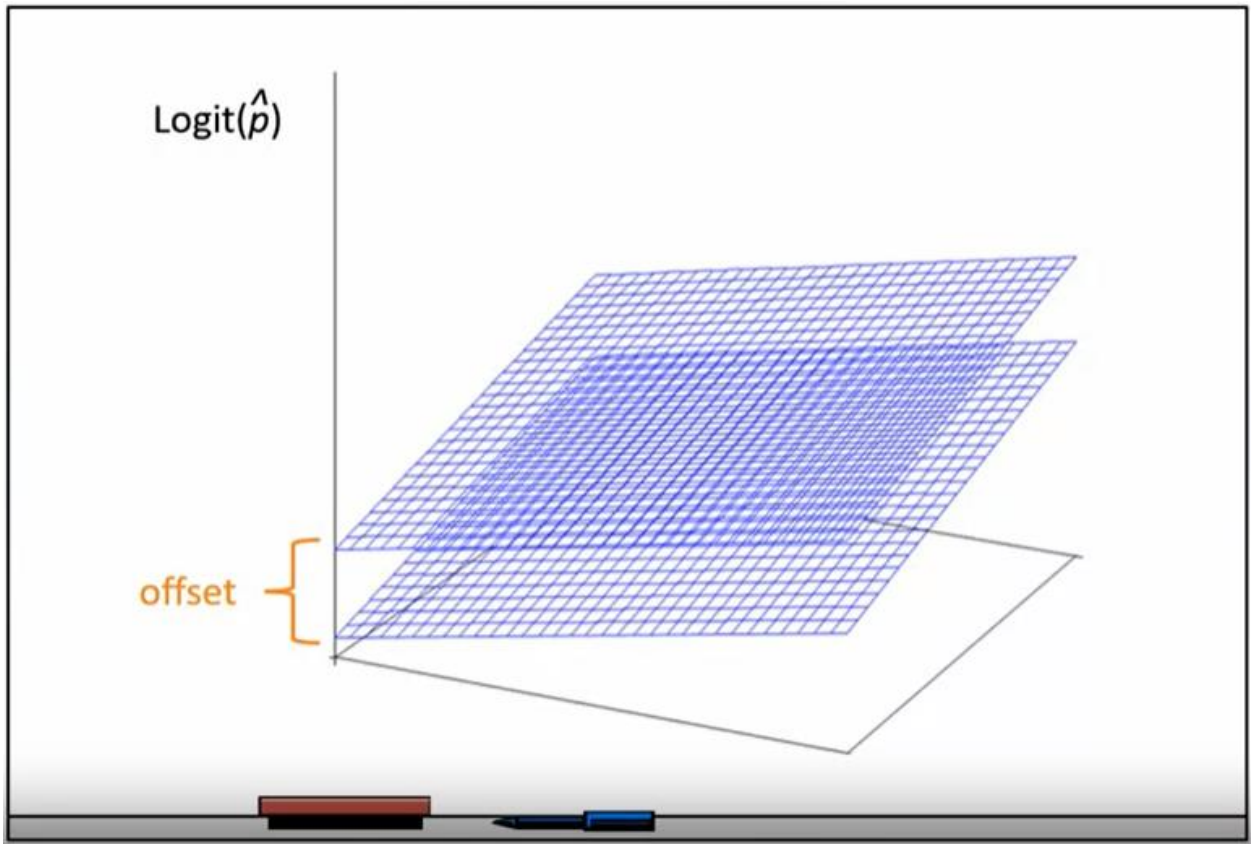

$\hat{p}$
$\hat{p}$
Biased
Corrected


Logit($\hat{p}$)
biased

Understanding the Offset



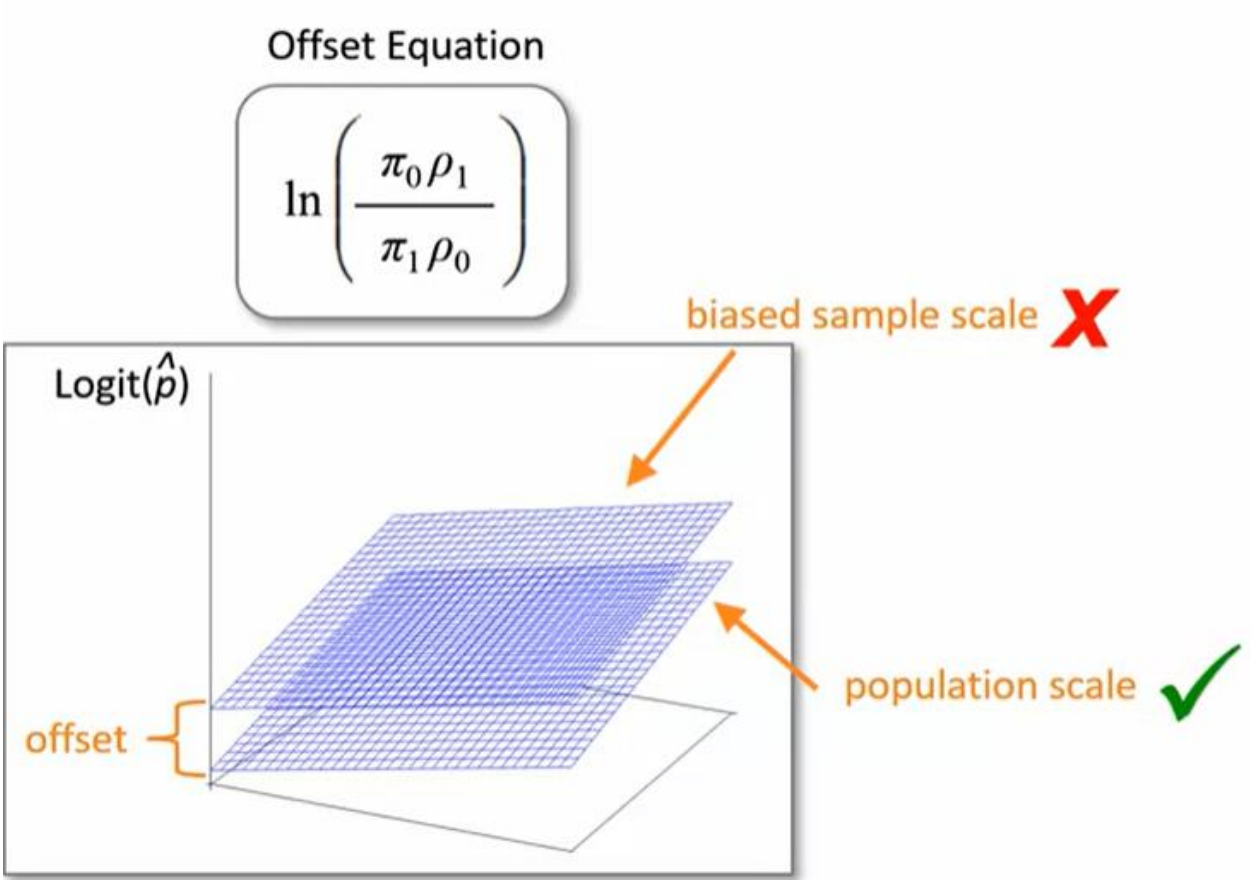



Logit($\hat{p}$)
offset

Offset Equation

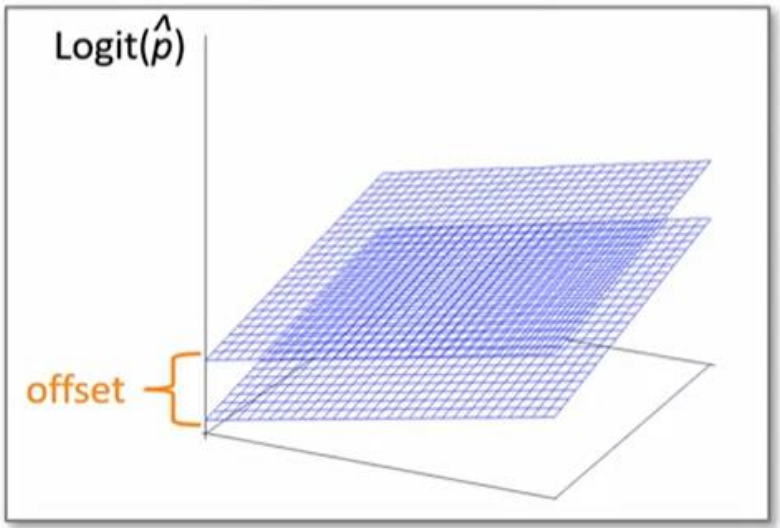
$$\ln\left(\frac{\pi_0 \rho_1}{\pi_1 \rho_0}\right)$$

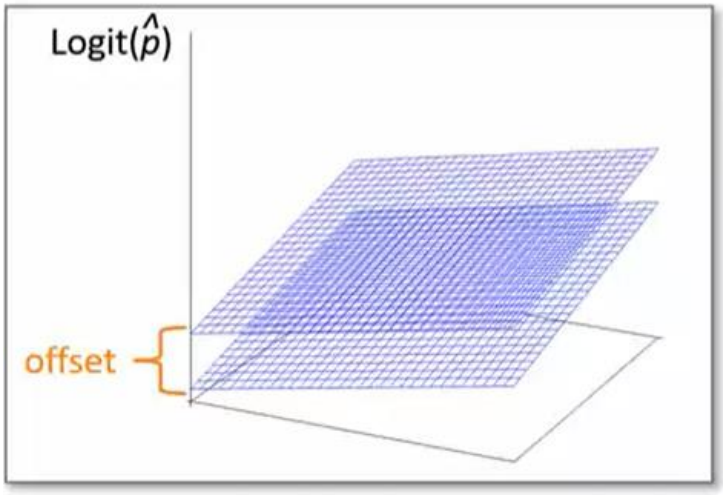
Offset Equation
π_0 = proportion of non-events in the population
π_1 = proportion of events in the population
ρ_1 = proportion of events in the sample
ρ_0 = proportion of non-events in the sample


Offset Equation
$\ln\left(\frac{\pi_0\rho_1}{\pi_1\rho_0}\right)$
biased sample scale
Logit($\hat{p}$)
population scale
offset

Offset Equation

$$\ln\left(\frac{\pi_0\rho_1}{\pi_1\rho_0}\right)$$



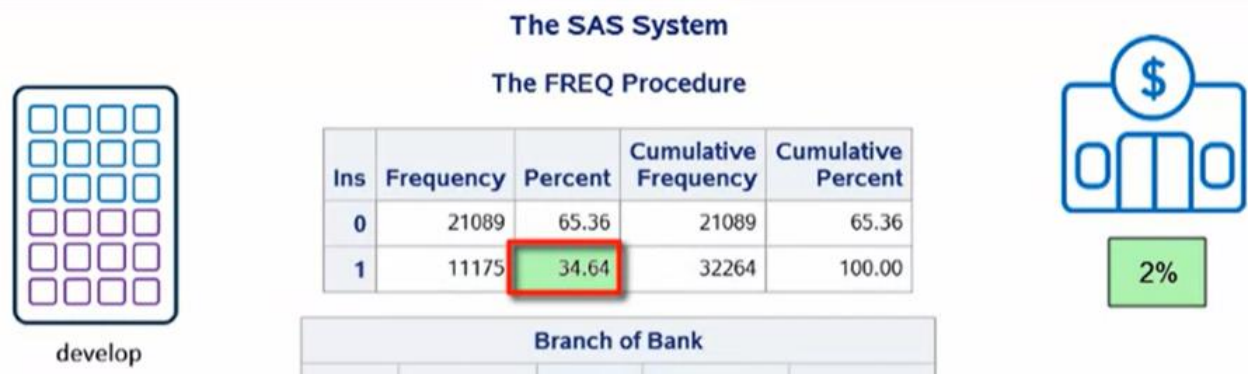


See Sampling Weights in the Resources section.

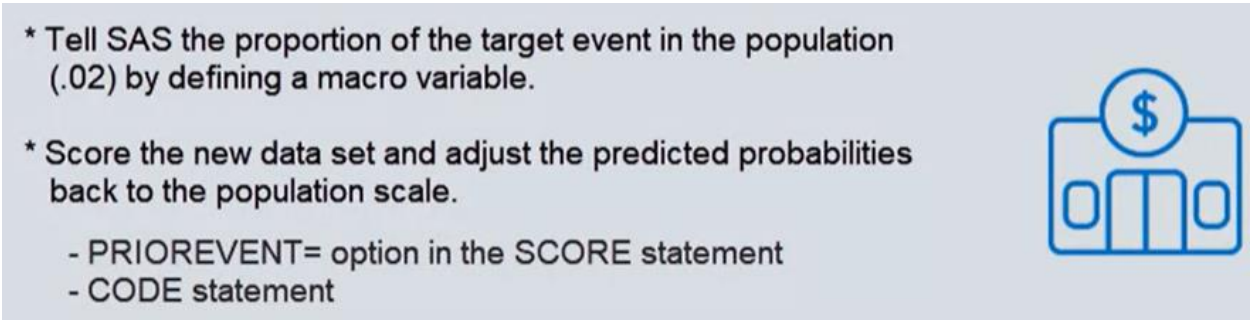
Demo Correcting for Oversampling

Predicted Probabilities from Scored Data Set

Obs	P_Ins1	DDA	DDABal	Dep	DepAmt	CashBk	Checks	Res
1	0.27023	1	56.29	2	955.51	0	1	U
2	0.32036	1	3292.17	2	961.60	0	1	U
3	0.29822	1	1723.86	2	2108.65	0	2	U
4	0.54208	0	0.00	0	0.00	0	0	U
5	0.26946	1	67.91	2	519.24	0	3	S
6	0.32228	1	2554.58	1	501.36	0	2	S
7	0.27038	1	0.00	2	2883.08	0	12	R
8	0.30399	1	2641.33	3	4521.61	0	8	S
9	0.54255	0	0.00	0	0.00	0	0	S
10	0.27956	1	52.22	1	75.59	0	0	R



Ins	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	21089	65.36	21089	65.36
1	11175	34.64	32264	100.00




```
/* ================================================== */
/* Lesson 2, Section 2: l2d3.sas
   Demonstration: Correcting for Oversampling
   [m642_2_f; derived from pmlr02d03.sas]            */
/* ================================================== */

/* Specify the prior probability to correct for oversampling.  */
%global pi1;
%let pi1=.02;

/* Correct predicted probabilities */

proc logistic data=work.train noprint;
   class res (param=ref ref='S');
   model ins(event='1')=dda ddabal dep depamt cashbk checks res;
   score data=pmlr.new out=work.scored4 priorevent=&pi1;
```

```
run;

title1 "Adjusted Predicted Probabilities from Scored Data Set";
proc print data=work.scored4(obs=10);
  var p_1 dda ddabal dep depamt cashbk checks res;
run;

title1 "Mean of Adjusted Predicted Probabilities from Scored Data Set";
proc means data=work.scored4 mean nolabels;
  var p_1;
run;
title1 ;
```

Adjusted Predicted Probabilities from Scored Data Set

Obs	P_1	DDA	DDABal	Dep	DepAmt	CashBk	Checks	Res
1	0.014060	1	56.29	2	955.51	0	1	U
2	0.017830	1	3292.17	2	961.60	0	1	U
3	0.016102	1	1723.86	2	2108.65	0	2	U
4	0.043602	0	0.00	0	0.00	0	0	U
5	0.014007	1	67.91	2	519.24	0	3	S
6	0.017985	1	2554.58	1	501.36	0	2	S
7	0.014071	1	0.00	2	2883.08	0	12	R
8	0.016543	1	2641.33	3	4521.61	0	8	S
9	0.043682	0	0.00	0	0.00	0	0	S
10	0.014724	1	52.22	1	75.59	0	0	R

Mean of Adjusted Predicted Probabilities from Scored Data Set

The MEANS Procedure

Analysis Variable : P_1
Mean
0.0249733

```
/* Correct probabilities in the Score Code */

proc logistic data=work.train noprint;
  class res (param=ref ref='S');
  model ins(event='1')=dda ddabal dep depamt cashbk checks res;
  /* File suffix "txt" is used so you can view the file */
  /* with a native text editor. SAS prefers "sas", but  */
  /* when specified as a filename, SAS does not care.   */
  code file="&PMLRfolder/pmlr_score_adj.txt";
run;

%global rho1;
proc SQL noprint;
  select mean(INS) into :rho1
  from work.train;
quit;

data new;
  set pmlr.new;
  off=log(((1-&pi1)*&rho1)/(&pi1*(1-&rho1)));
run;

data work.scored5;
  set work.new;
  %include "&PMLRfolder/pmlr_score_adj.txt";
  eta=log(p_ins1/p_ins0) - off;
  prob=1/(1+exp(-eta));
run;
```

```
title1 "Adjusted Predicted Probabilities from Scored Data Set";
proc print data=scored5(obs=10);
  var prob dda ddabal dep depamt cashbk checks res;
run;
title1 ;
```

Adjusted Predicted Probabilities from Scored Data Set

Obs	prob	DDA	DDABal	Dep	DepAmt	CashBk	Checks	Res
1	0.014060	1	56.29	2	955.51	0	1	U
2	0.017830	1	3292.17	2	961.60	0	1	U
3	0.016102	1	1723.86	2	2108.65	0	2	U
4	0.043602	0	0.00	0	0.00	0	0	U
5	0.014007	1	67.91	2	519.24	0	3	S
6	0.017985	1	2554.58	1	501.36	0	2	S
7	0.014071	1	0.00	2	2883.08	0	12	R
8	0.016543	1	2641.33	3	4521.61	0	8	S
9	0.043682	0	0.00	0	0.00	0	0	S
10	0.014724	1	52.22	1	75.59	0	0	R

```
/* Run this code before doing practice l2p1 */


/* ================================================== */
/* Lesson 1, Practice 1
   Practice: Exploring the Veterans' Organization Data
   Used in the Practices                              */
/* ================================================== */

data pmlr.pva(drop=control_number
              MONTHS_SINCE_LAST_PROM_RESP
              FILE_AVG_GIFT
              FILE_CARD_GIFT);
```

```
  set pmlr.pva_raw_data;
  STATUS_FL=RECENCY_STATUS_96NK in("F","L");
  STATUS_ES=RECENCY_STATUS_96NK in("E","S");
  home01=(HOME_OWNER="H");
  nses1=(SES="1");
  nses3=(SES="3");
  nses4=(SES="4");
  nses_=(SES="?");
  nurbr=(URBANICITY="R");
  nurbu=(URBANICITY="U");
  nurbs=(URBANICITY="S");
  nurbt=(URBANICITY="T");
  nurb_=(URBANICITY="?");
run;

proc contents data=pmlr.pva;
run;

proc means data=pmlr.pva mean nmiss max min;
  var _numeric_;
run;

proc freq data=pmlr.pva nlevels;
  tables _character_;
run;



/* ================================================== */
/* Lesson 1, Practice 2
```

```
  Practice: Splitting the Data                    */
/* ================================================== */

proc sort data=pmlr.pva out=work.pva_sort;
  by target_b;
run;

proc surveyselect noprint data=work.pva_sort
              samprate=0.5 out=pva_sample seed=27513
              outall stratumseed=restore;
  strata target_b;
run;

data pmlr.pva_train(drop=selected SelectionProb SamplingWeight)
    pmlr.pva_valid(drop=selected SelectionProb SamplingWeight);
  set work.pva_sample;
  if selected then output pmlr.pva_train;
  else output pmlr.pva_valid;
run;
```

The CONTENTS Procedure

Data Set Name	PMLR.PVA	Observations	19372
Member Type	DATA	Variables	58
Engine	V9	Indexes	0
Created	09/16/2021 23:22:18	Observation Length	432
Last Modified	09/16/2021 23:22:18	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	65
First Data Page	1
Max Obs per Page	303
Obs in First Data Page	281
Number of Data Set Repairs	0
Filename	/home/u58304328/EPMLR51/data/pva.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	10886327280
Access Permission	rw-r--r--
Owner Name	u58304328
File Size	8MB
File Size (bytes)	8650752

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
42	CARD_PROM_12	Num	8
8	CLUSTER_CODE	Char	2
4	DONOR_AGE	Num	8
10	DONOR_GENDER	Char	3
26	FREQUENCY_STATUS_97NK	Num	8
9	HOME_OWNER	Char	3
11	INCOME_GROUP	Num	8
5	IN_HOUSE	Num	8
41	LAST_GIFT_AMT	Num	8
37	LIFETIME_AVG_GIFT_AMT	Num	8
33	LIFETIME_CARD_PROM	Num	8
35	LIFETIME_GIFT_AMOUNT	Num	8
36	LIFETIME_GIFT_COUNT	Num	8
38	LIFETIME_GIFT_RANGE	Num	8
39	LIFETIME_MAX_GIFT_AMT	Num	8
40	LIFETIME_MIN_GIFT_AMT	Num	8
34	LIFETIME_PROM	Num	8
16	MEDIAN_HOME_VALUE	Num	8
17	MEDIAN_HOUSEHOLD_INCOME	Num	8
45	MONTHS_SINCE_FIRST_GIFT	Num	8
44	MONTHS_SINCE_LAST_GIFT	Num	8
3	MONTHS_SINCE_ORIGIN	Num	8
14	MOR_HIT_RATE	Num	8
43	NUMBER_PROM_12	Num	8
13	OVERLAY_SOURCE	Char	1
19	PCT_MALE_MILITARY	Num	8

20	PCT_MALE_VETERANS	Num	8
18	PCT_OWNER_OCCUPIED	Num	8
21	PCT_VIETNAM_VETERANS	Num	8
22	PCT_WWII_VETERANS	Num	8
23	PEP_STAR	Num	8
46	PER_CAPITA_INCOME	Num	8
12	PUBLISHED_PHONE	Num	8
25	RECENCY_STATUS_96NK	Char	5
30	RECENT_AVG_CARD_GIFT_AMT	Num	8
28	RECENT_AVG_GIFT_AMT	Num	8
32	RECENT_CARD_RESPONSE_COUNT	Num	8
29	RECENT_CARD_RESPONSE_PROP	Num	8
31	RECENT_RESPONSE_COUNT	Num	8
27	RECENT_RESPONSE_PROP	Num	8
24	RECENT_STAR_STATUS	Num	8
7	SES	Char	4
48	STATUS_ES	Num	8
47	STATUS_FL	Num	8
1	TARGET_B	Num	8
2	TARGET_D	Num	8
6	URBANICITY	Char	4
15	WEALTH_RATING	Num	8
49	home01	Num	8
50	nses1	Num	8
51	nses3	Num	8
52	nses4	Num	8
53	nses_	Num	8

The MEANS Procedure

Variable	Mean	N Miss	Maximum	Minimum
TARGET_B	0.2500000	0	1.0000000	0
TARGET_D	15.6243444	14529	200.0000000	1.0000000
MONTHS_SINCE_ORIGIN	73.4099732	0	137.0000000	5.0000000
DONOR_AGE	58.9190506	4795	87.0000000	0
IN_HOUSE	0.0731984	0	1.0000000	0
INCOME_GROUP	3.9075434	4392	7.0000000	1.0000000
PUBLISHED_PHONE	0.4977287	0	1.0000000	0
MOR_HIT_RATE	3.3616560	0	241.0000000	0
WEALTH_RATING	5.0053967	8810	9.0000000	0
MEDIAN_HOME_VALUE	1079.87	0	6000.00	0
MEDIAN_HOUSEHOLD_INCOME	341.9702147	0	1500.00	0
PCT_OWNER_OCCUPIED	69.6989986	0	99.0000000	0
PCT_MALE_MILITARY	1.0290109	0	97.0000000	0
PCT_MALE_VETERANS	30.5739211	0	99.0000000	0
PCT_VIETNAM_VETERANS	29.6032934	0	99.0000000	0
PCT_WWII_VETERANS	32.8524675	0	99.0000000	0
PEP_STAR	0.5044394	0	1.0000000	0
RECENT_STAR_STATUS	0.9311377	0	22.0000000	0
FREQUENCY_STATUS_97NK	1.9839975	0	4.0000000	1.0000000
RECENT_RESPONSE_PROP	0.1901275	0	1.0000000	0
RECENT_AVG_GIFT_AMT	15.3653959	0	260.0000000	0
RECENT_CARD_RESPONSE_PROP	0.2308077	0	1.0000000	0
RECENT_AVG_CARD_GIFT_AMT	11.6854703	0	300.0000000	0
RECENT_RESPONSE_COUNT	3.0431034	0	16.0000000	0
RECENT_CARD_RESPONSE_COUNT	1.7305389	0	9.0000000	0
LIFETIME_CARD_PROM	18.6680776	0	56.0000000	2.0000000
LIFETIME_PROM	47.5705141	0	194.0000000	5.0000000
LIFETIME_GIFT_AMOUNT	104.4257165	0	3775.00	15.0000000
LIFETIME_GIFT_COUNT	9.9797646	0	95.0000000	1.0000000
LIFETIME_AVG_GIFT_AMT	12.8583383	0	450.0000000	1.3600000
LIFETIME_GIFT_RANGE	11.5878758	0	997.0000000	0
LIFETIME_MAX_GIFT_AMT	19.2088081	0	1000.00	5.0000000
LIFETIME_MIN_GIFT_AMT	7.6209323	0	450.0000000	0
LAST_GIFT_AMT	16.5841988	0	450.0000000	0
CARD_PROM_12	5.3671278	0	17.0000000	0
NUMBER_PROM_12	12.9018687	0	64.0000000	2.0000000
MONTHS_SINCE_LAST_GIFT	18.1911522	0	27.0000000	4.0000000
MONTHS_SINCE_FIRST_GIFT	69.4820875	0	260.0000000	15.0000000
PER_CAPITA_INCOME	15857.33	0	174523.00	0
STATUS_FL	0.0833161	0	1.0000000	0

The FREQ Procedure

Number of Variable Levels	
Variable	Levels
URBANICITY	6
SES	5
CLUSTER_CODE	54
HOME_OWNER	2
DONOR_GENDER	4
OVERLAY_SOURCE	4
RECENCY_STATUS_96NK	6

URBANICITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	454	2.34	454	2.34
C	4022	20.76	4476	23.11
R	4005	20.67	8481	43.78
S	4491	23.18	12972	66.96
T	3944	20.36	16916	87.32
U	2456	12.68	19372	100.00

SES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5924	30.58	5924	30.58
2	9284	47.92	15208	78.51
3	3323	17.15	18531	95.66
4	387	2.00	18918	97.66
?	454	2.34	19372	100.00

CLUSTER_CODE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	454	2.34	454	2.34
01	239	1.23	693	3.58
02	380	1.96	1073	5.54
03	300	1.55	1373	7.09
04	113	0.58	1486	7.67
05	199	1.03	1685	8.70
06	123	0.63	1808	9.33
07	184	0.95	1992	10.28
08	378	1.95	2370	12.23
09	153	0.79	2523	13.02
10	387	2.00	2910	15.02
11	484	2.50	3394	17.52
12	631	3.26	4025	20.78
13	579	2.99	4604	23.77
14	454	2.34	5058	26.11
15	223	1.15	5281	27.26
16	384	1.98	5665	29.24
17	349	1.80	6014	31.04
18	619	3.20	6633	34.24
19	98	0.51	6731	34.75
20	317	1.64	7048	36.38
21	353	1.82	7401	38.20
22	251	1.30	7652	39.50
23	293	1.51	7945	41.01
24	795	4.10	8740	45.12
25	273	1.41	9013	46.53

26	202	1.04	9215	47.57
27	666	3.44	9881	51.01
28	343	1.77	10224	52.78
29	170	0.88	10394	53.65
30	519	2.68	10913	56.33
31	249	1.29	11162	57.62
32	152	0.78	11314	58.40
33	109	0.56	11423	58.97
34	284	1.47	11707	60.43
35	727	3.75	12434	64.19
36	716	3.70	13150	67.88
37	204	1.05	13354	68.93
38	240	1.24	13594	70.17
39	512	2.64	14106	72.82
40	830	4.28	14936	77.10
41	431	2.22	15367	79.33
42	284	1.47	15651	80.79
43	468	2.42	16119	83.21
44	383	1.98	16502	85.18
45	482	2.49	16984	87.67
46	369	1.90	17353	89.58
47	185	0.95	17538	90.53
48	180	0.93	17718	91.46
49	675	3.48	18393	94.95
50	156	0.81	18549	95.75
51	460	2.37	19009	98.13
52	60	0.31	19069	98.44
53	303	1.56	19372	100.00

HOME_OWNER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H	10606	54.75	10606	54.75
U	8766	45.25	19372	100.00

DONOR_GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	1	0.01	1	0.01
F	10401	53.69	10402	53.70
M	7953	41.05	18355	94.75
U	1017	5.25	19372	100.00

OVERLAY_SOURCE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
B	8732	45.08	8732	45.08
M	1480	7.64	10212	52.72
N	4392	22.67	14604	75.39
P	4768	24.61	19372	100.00

RECENCY_STATUS_96NK	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	11918	61.52	11918	61.52
E	427	2.20	12345	63.73
F	1521	7.85	13866	71.58
L	93	0.48	13959	72.06
N	1192	6.15	15151	78.21
S	4221	21.79	19372	100.00

/* Solution for l2p1 */

/* step 2 */

```
%global ex_pi1;
%let ex_pi1=0.05;
```

```
/* step 3 */

title1 "Logistic Regression Model of the Veterans' Organization Data";
proc logistic data=pmlr.pva_train plots(only)=
          (effect(clband x=(pep_star recent_avg_gift_amt
          frequency_status_97nk)) oddsratio (type=horizontalstat));
  class pep_star (param=ref ref='0');
  model target_b(event='1')=pep_star recent_avg_gift_amt
             frequency_status_97nk / clodds=pl;
  effectplot slicefit(sliceby=pep_star x=recent_avg_gift_amt) / noobs;
  effectplot slicefit(sliceby=pep_star x=frequency_status_97nk) / noobs;
  score data=pmlr.pva_train out=work.scopva_train priorevent=&ex_pi1;
run;



/* step 5 */

title1 "Adjusted Predicted Probabilities of the Veteran's Organization Data";
proc print data=work.scopva_train(obs=10);
  var p_1 pep_star recent_avg_gift_amt frequency_status_97nk;
run;
title;
```

Logistic Regression Model of the Veterans' Organization Data

The LOGISTIC Procedure

Model Information	
Data Set	PMLR.PVA_TRAIN
Response Variable	TARGET_B
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	9687
Number of Observations Used	9687

Response Profile		
Ordered Value	TARGET_B	Total Frequency
1	0	7265
2	1	2422

Probability modeled is TARGET_B=1.

Class Level Information		
Class	Value	Design Variables
PEP_STAR	0	0
	1	1

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	10897.230	10663.061
SC	10904.409	10691.776
-2 Log L	10895.230	10655.061

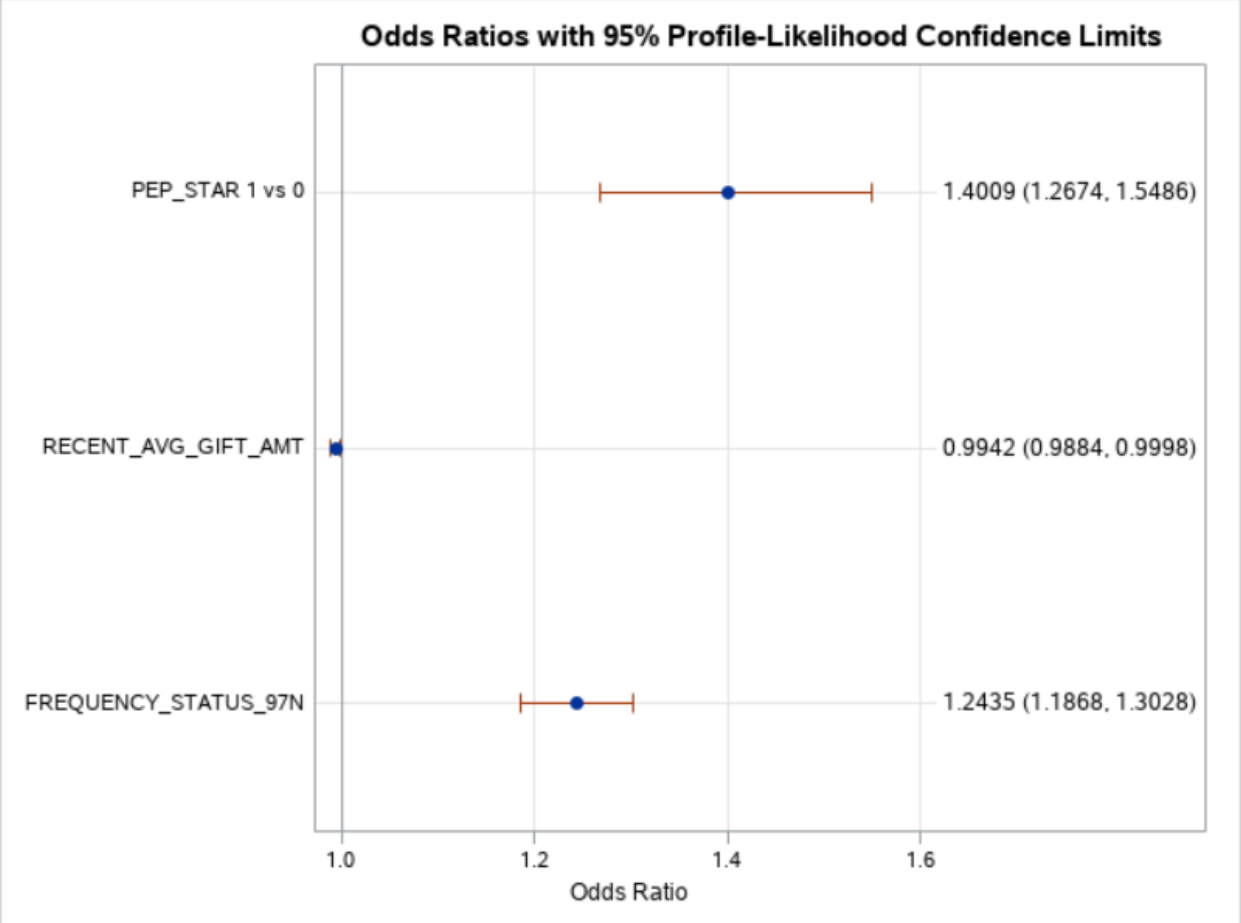
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	240.1690	3	<.0001
Score	242.9486	3	<.0001
Wald	237.2875	3	<.0001

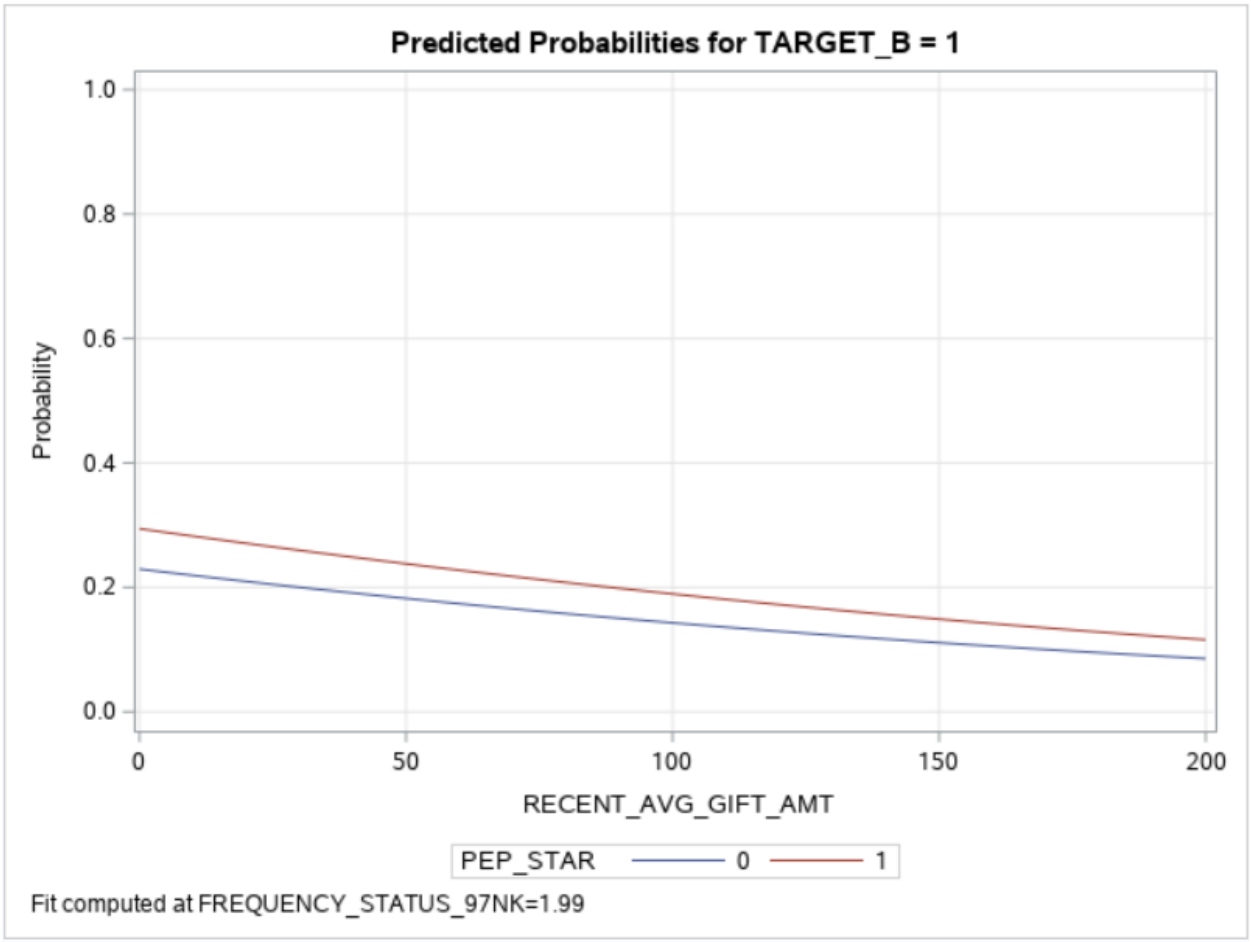
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
PEP_STAR	1	43.4902	<.0001
RECENT_AVG_GIFT_AMT	1	3.9559	0.0467
FREQUENCY_STATUS_97N	1	83.8209	<.0001

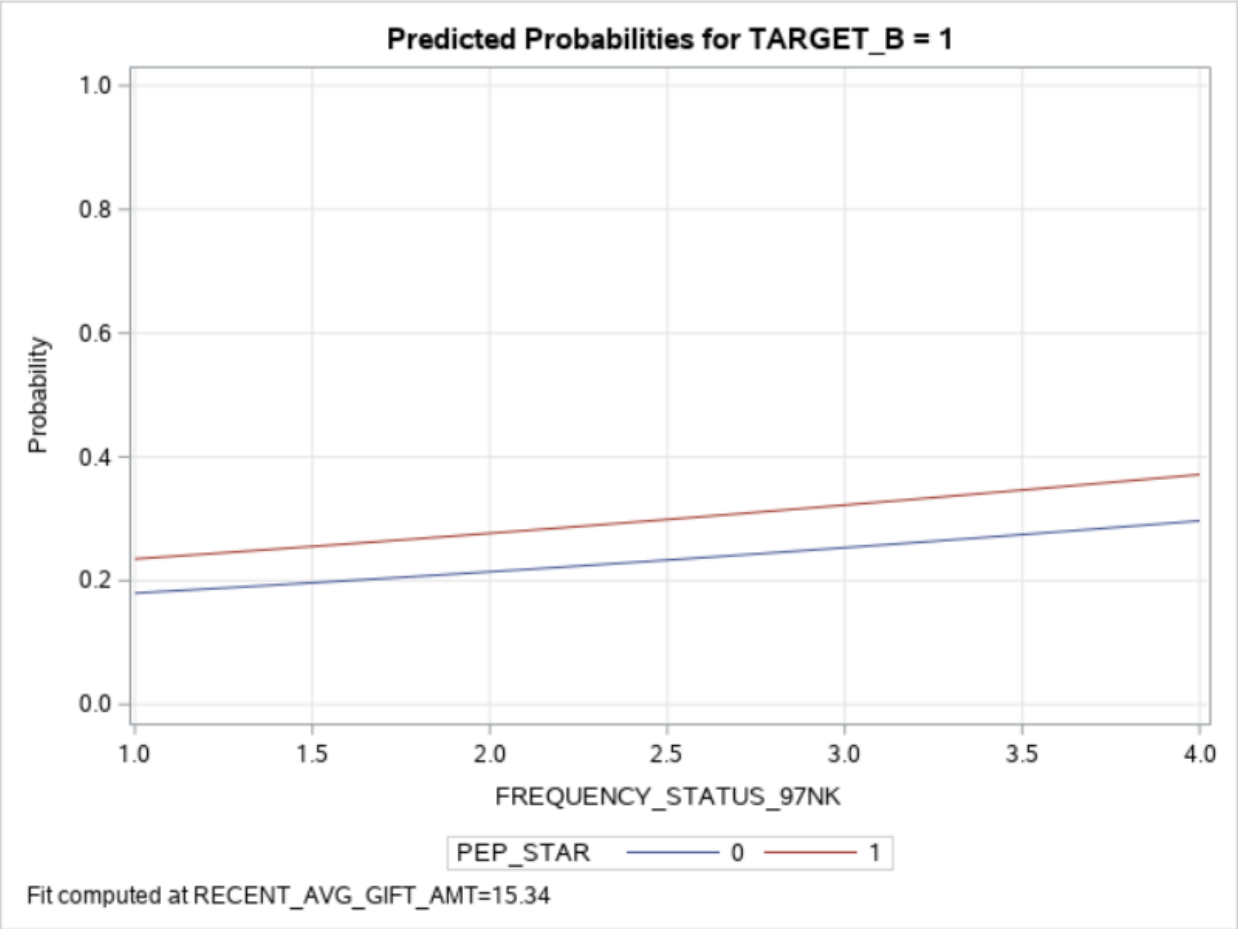
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.6454	0.0831	392.4480	<.0001
PEP_STAR	1	1	0.3371	0.0511	43.4902	<.0001
RECENT_AVG_GIFT_AMT		1	-0.00579	0.00291	3.9559	0.0467
FREQUENCY_STATUS_97N		1	0.2179	0.0238	83.8209	<.0001

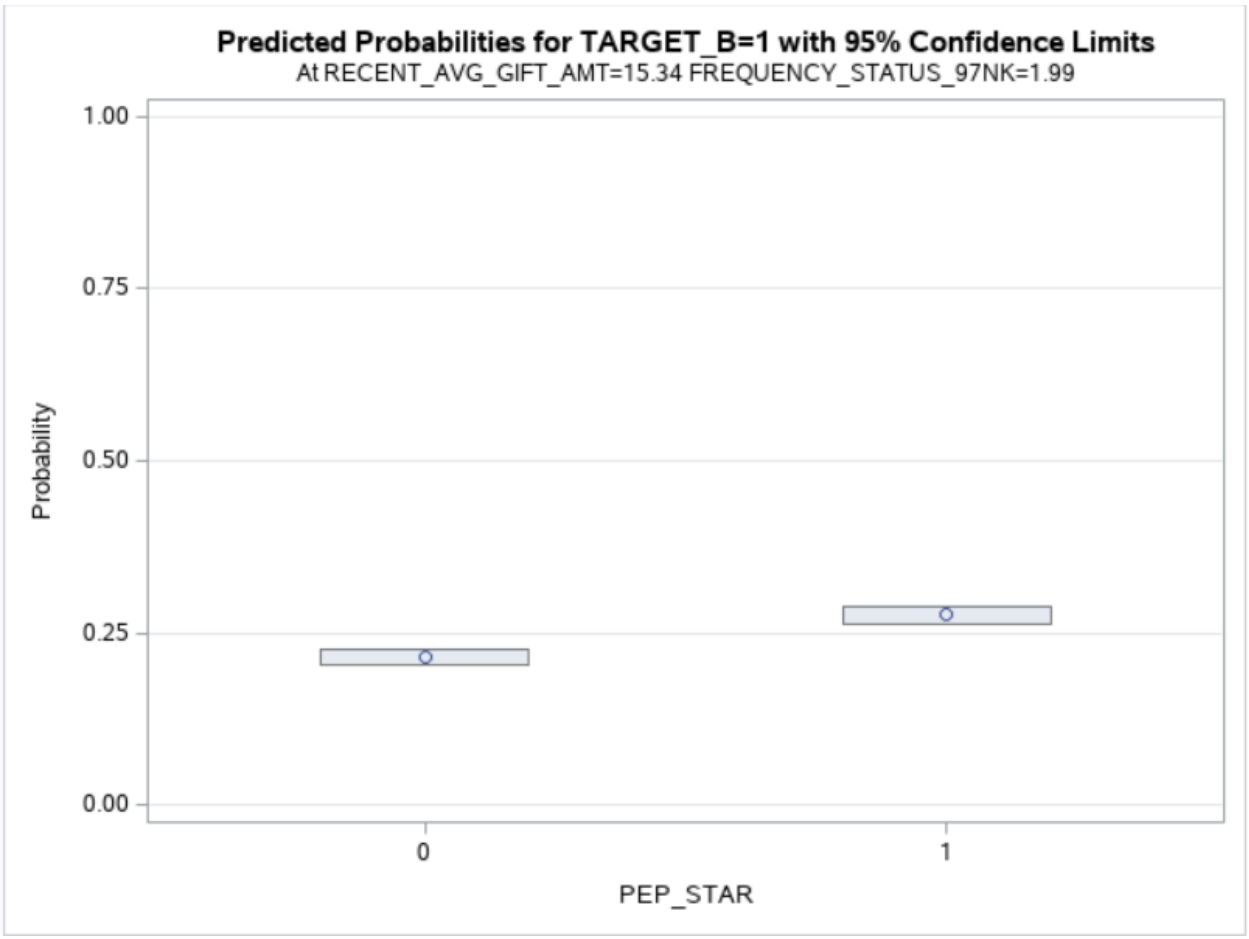
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	59.9	Somers' D	0.208
Percent Discordant	39.0	Gamma	0.211
Percent Tied	1.1	Tau-a	0.078
Pairs	17595830	c	0.604

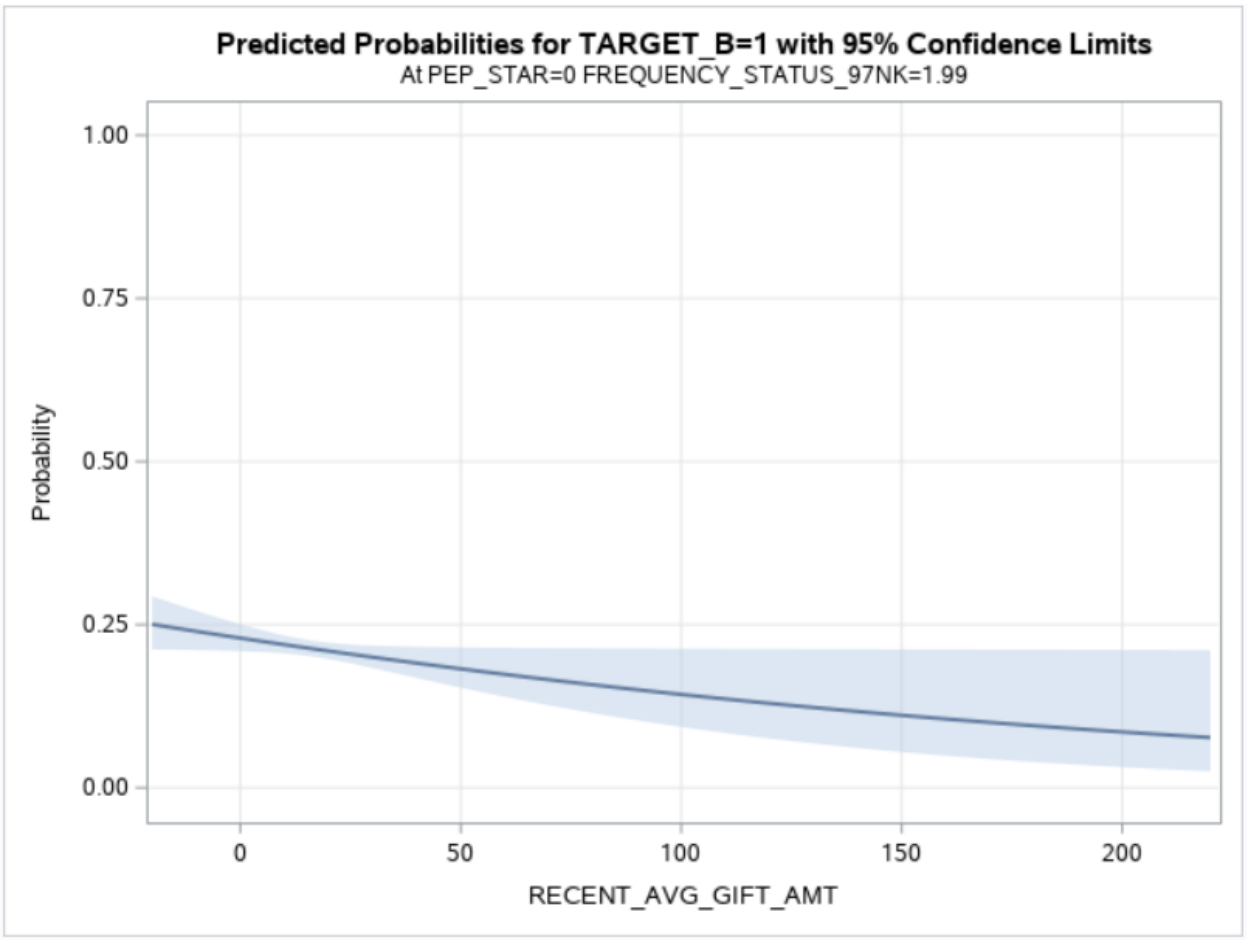
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
PEP_STAR 1 vs 0	1.0000	1.401	1.267	1.549
RECENT_AVG_GIFT_AMT	1.0000	0.994	0.988	1.000
FREQUENCY_STATUS_97N	1.0000	1.243	1.187	1.303

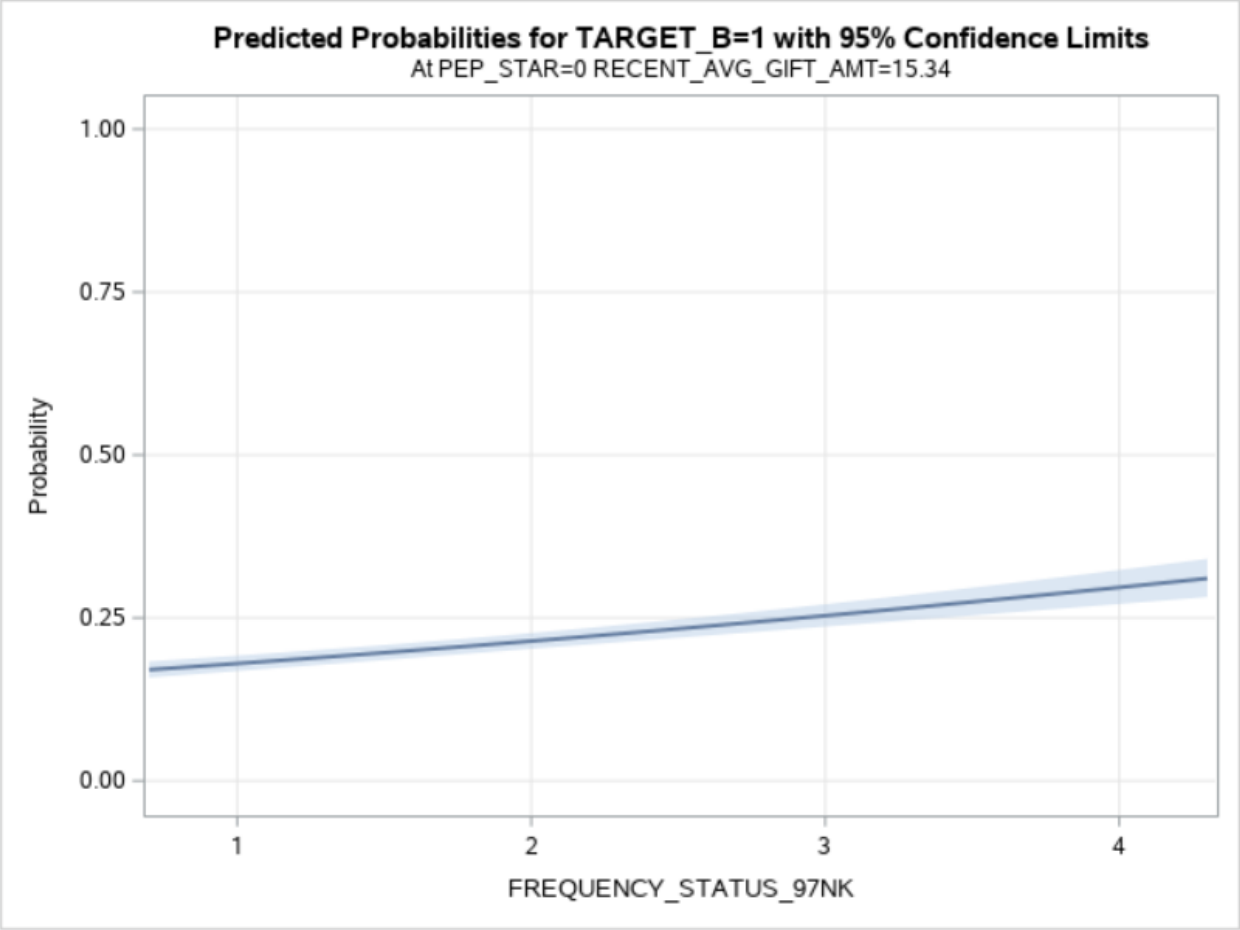



Predicted Probabilities for TARGET_B = 1
1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
50
100
150
200
RECENT_AVG_GIFT_AMT
PEP_STAR
0
1
Fit computed at FREQUENCY_STATUS_97NK=1.99


Predicted Probabilities for TARGET_B = 1
Probability
1.0
0.8
0.6
0.4
0.2
0.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
FREQUENCY_STATUS_97NK
PEP_STAR
0
1
Fit computed at RECENT_AVG_GIFT_AMT=15.34


Predicted Probabilities for TARGET_B=1 with 95% Confidence Limits
At RECENT_AVG_GIFT_AMT=15.34 FREQUENCY_STATUS_97NK=1.99
1.00
0.75
0.50
0.25
0.00
Probability
0
1
PEP_STAR


Predicted Probabilities for TARGET_B=1 with 95% Confidence Limits
At PEP_STAR=0 FREQUENCY_STATUS_97NK=1.99
Probability
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
RECENT_AVG_GIFT_AMT



Adjusted Predicted Probabilities of the Veteran's Organization Data

Obs	P_1	PEP_STAR	RECENT_AVG_GIFT_AMT	FREQUENCY_STATUS_97NK
1	0.046390	1	15.00	1
2	0.033094	0	17.50	1
3	0.064890	0	8.33	4
4	0.090167	1	5.00	4
5	0.059152	1	8.33	2
6	0.058117	1	11.57	2
7	0.046941	1	12.86	1
8	0.031733	0	25.00	1
9	0.045126	1	20.00	1
10	0.032091	0	23.00	1

Practice: Fitting a Logistic Regression Model

Question 1

For the veterans' organization project, fit a logistic regression model to the **pmlr.pva_train** data set and use ODS statistical graphics to display the results.

Reminder: If you started a new SAS session, you must run **setup.sas** to define the **pmlr** library before you do this practice.

Step 1: Open **l2p01_runFirst.sas** from the **practices** folder and run the code. You can add to this program or open a new editor to continue the practice.

Step 2: Create a global macro variable named **ex_pi1** that stores $\pi 1$, the proportion of responders in the population. Note: To find the proportion of responders in the population, review the **pva_raw_data** data set description.

What value did you assign to **ex_pi1**?

The proportion of events in the population is estimated at 0.05.

For the solution code, open **l2p1_s.sas** from the **practices/solutions** folder and see Step 2.

Question 2

Step 3: Add a PROC LOGISTIC step that does the following:

- fits a logistic regression model with **Target_B** as the target variable, and **Pep_Star**, **Recent_Avg_Gift_Amt**, and **Frequency_Status_97NK** as the input variables
- models the probability that the target variable equals 1, and requests 95% profile likelihood confidence intervals for the odds ratio
- uses the SCORE statement to create a temporary data set called scopva_train, which contains the data from the pva_train data set and the predicted probability of the event, correcting for oversampling
- creates effect plots with confidence bands for the three input variables
- creates effect plots of **Recent_Avg_Gift_Amt** by **Pep_Star** and **Frequency_Status_97NK** by **PEP_STAR**
- creates an odds ratio plot with a horizontal orientation and displays the statistics
- uses the ONLY global plot option to suppress the default plots

Note: To avoid a warning in the log about the suppression of plots that have more than 5000 observations, you can add the following MAXPOINTS= option to the PROC LOGISTIC statement: **plots(maxpoints=none only)** This change is optional. Omitting the MAXPOINTS= option does not affect the results of the practices in this course.

Submit the code and interpret the following metrics:

- the c statistic
- the odds ratio for **Pep_Star**
- the effect plot for **Recent_Avg_Gift_Amt** by **Pep_Star**
- the effect plot for **Recent_Avg_Gift_Amt**

Note: This is a free response question and all attempts are marked correct. Type your response and compare your answer to the answer provided.

Based on the results, you interpret the specified metrics as follows:

- The *c* statistic is shown in the Association of Predicted Probabilities and Observed Responses table. The *c* statistic, 0.604, can be interpreted as the probability of a customer who donated to the national veterans' organization having a higher predicted probability (of donating to the organization) compared to a customer who did not donate.
- The odds ratio for **Pep_Star** is shown in the Odds Ratio Estimates and Profile-Likelihood Confidence Intervals table. This odds ratio indicates that customers who are consecutive donors have 1.40 times the odds of responding to a solicitation compared to customers who are not consecutive donors.
- The effect plot for **Recent_Avg_Gift_Amt** by **Pep_Star** shows a negative relationship between the average donation amount in response to promotions since June 1994 and the predicted probabilities of responding to the solicitation in June of 1997. The consecutive donors have the higher probabilities across all the values of **Recent_Avg_Gift_Amt**. The consecutive donors have the highest probabilities across all values of **Recent_Avg_Gift_Amt**.
- The effect plot for **Recent_Avg_Gift_Amt** shows a negative relationship between the average donation amount in response to promotions since June 1994 and the predicted probabilities of responding to the solicitation in June 1997. The highest confidence intervals correspond to the largest values of **Recent_Avg_Gift_Amt**.

For the solution code, open **l2p1_s.sas** in the **practices/solutions** folder and see Step 3.

```
data work.pva(drop=CONTROL_NUMBER MONTHS_SINCE_LAST_PROM_RESP
          FILE_AVG_GIFT FILE_CARD_GIFT);
  set pmlr.pva_raw_data;
run;

data work.pva;
  set work.pva;
  STATUS_FL=RECENCY_STATUS_96NK in("F","L");
  STATUS_ES=RECENCY_STATUS_96NK in("E","S");
run;

proc logistic data=work.pva plots(only)=(effect(clband
          x=(lifetime_card_prom recent_response_prop
            months_since_last_gift
            recent_avg_gift_amt status_es))
            oddsratio (type=horizontalstat))
            namelen=25;
  model target_b(event='1')= lifetime_card_prom
      recent_response_prop months_since_last_gift
      recent_avg_gift_amt status_es / clodds=pl stb;
  units lifetime_card_prom=10 months_since_last_gift=6
      recent_avg_gift_amt=25 / default=1;
run;
```

The LOGISTIC Procedure

Model Information	
Data Set	WORK.PVA
Response Variable	TARGET_B
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	19372
Number of Observations Used	19372

Response Profile		
Ordered Value	TARGET_B	Total Frequency
1	0	14529
2	1	4843

Probability modeled is TARGET_B=1.

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

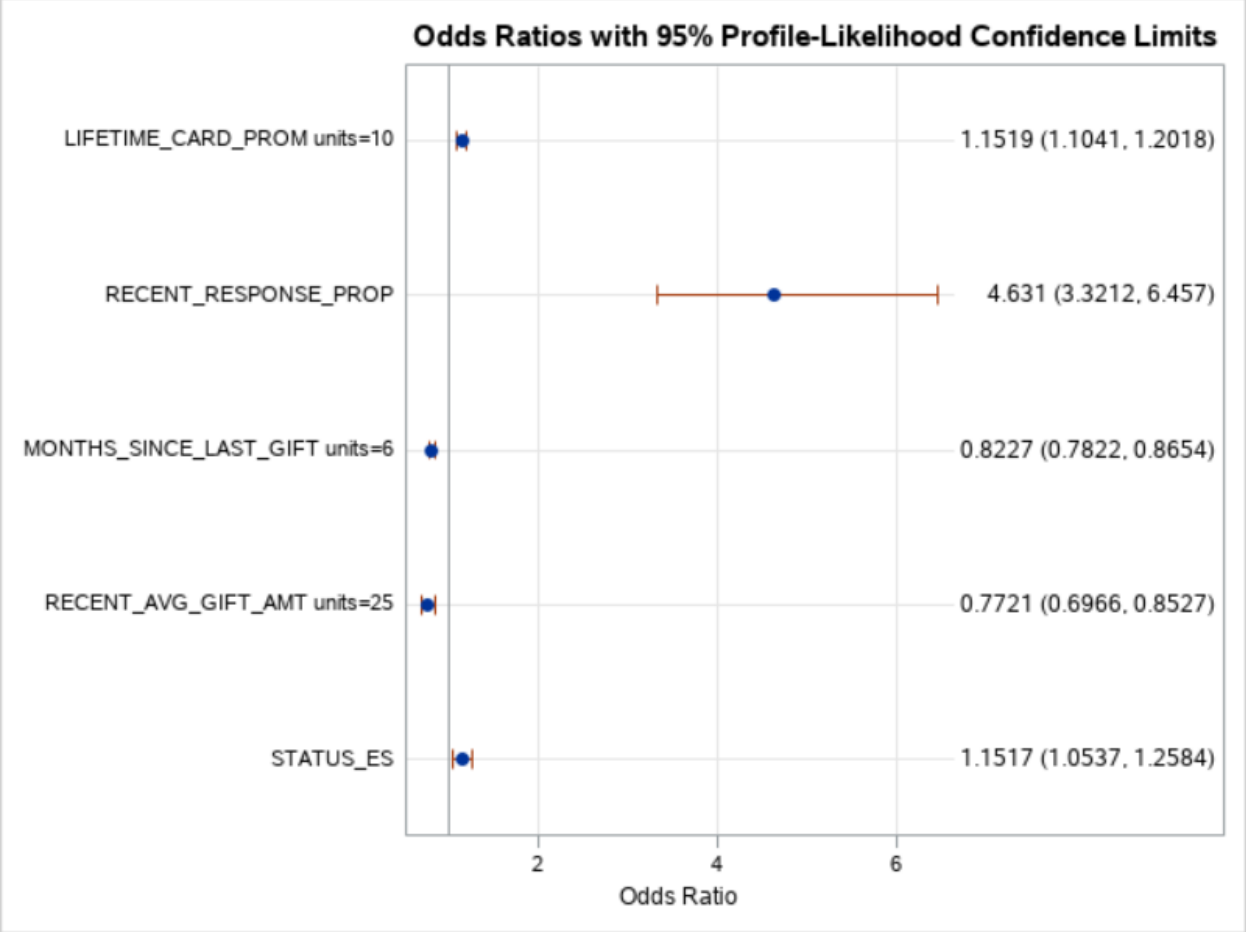
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	21789.113	21345.336
SC	21796.984	21392.566
-2 Log L	21787.113	21333.336

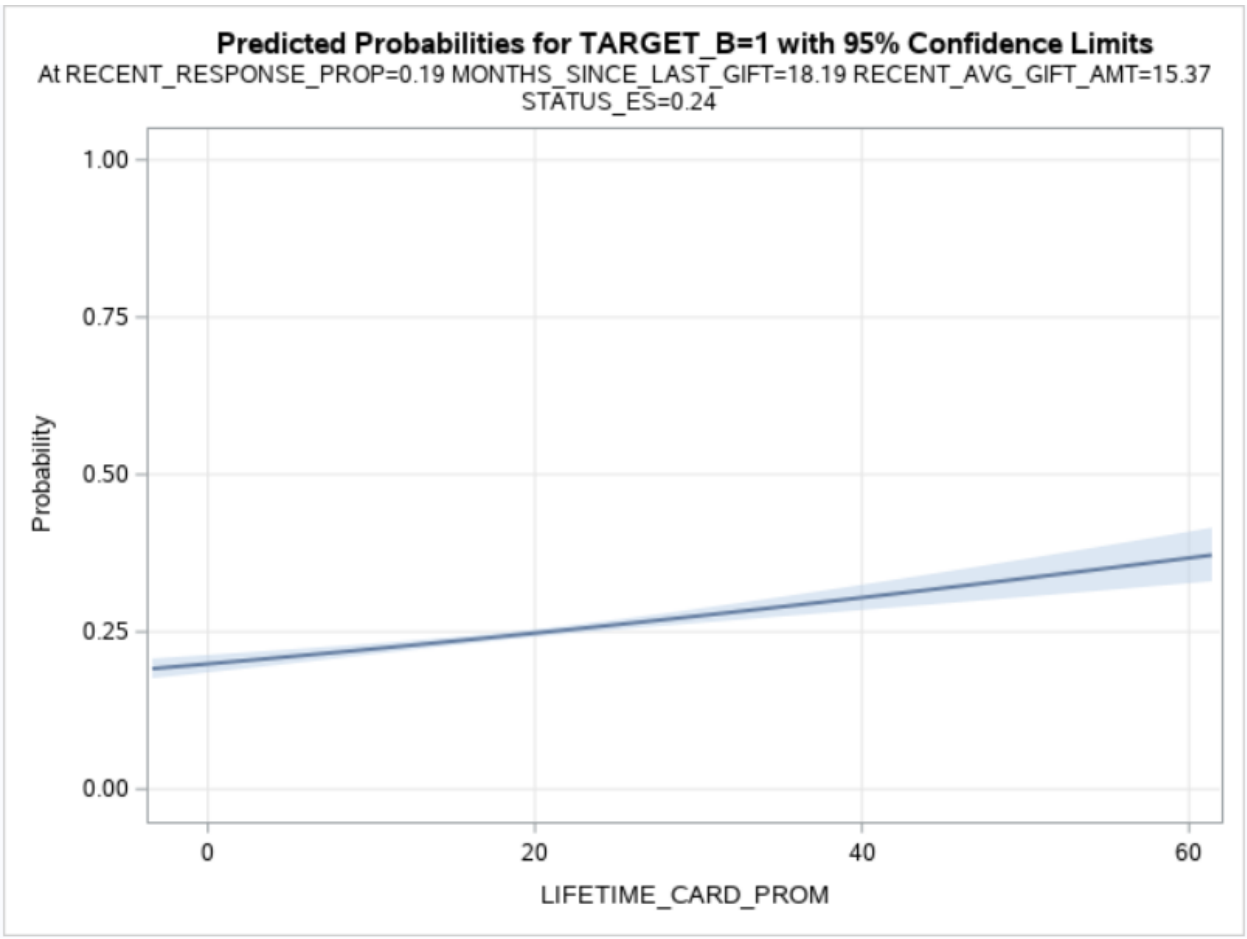
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	453.7764	5	<.0001
Score	461.6635	5	<.0001
Wald	446.9425	5	<.0001

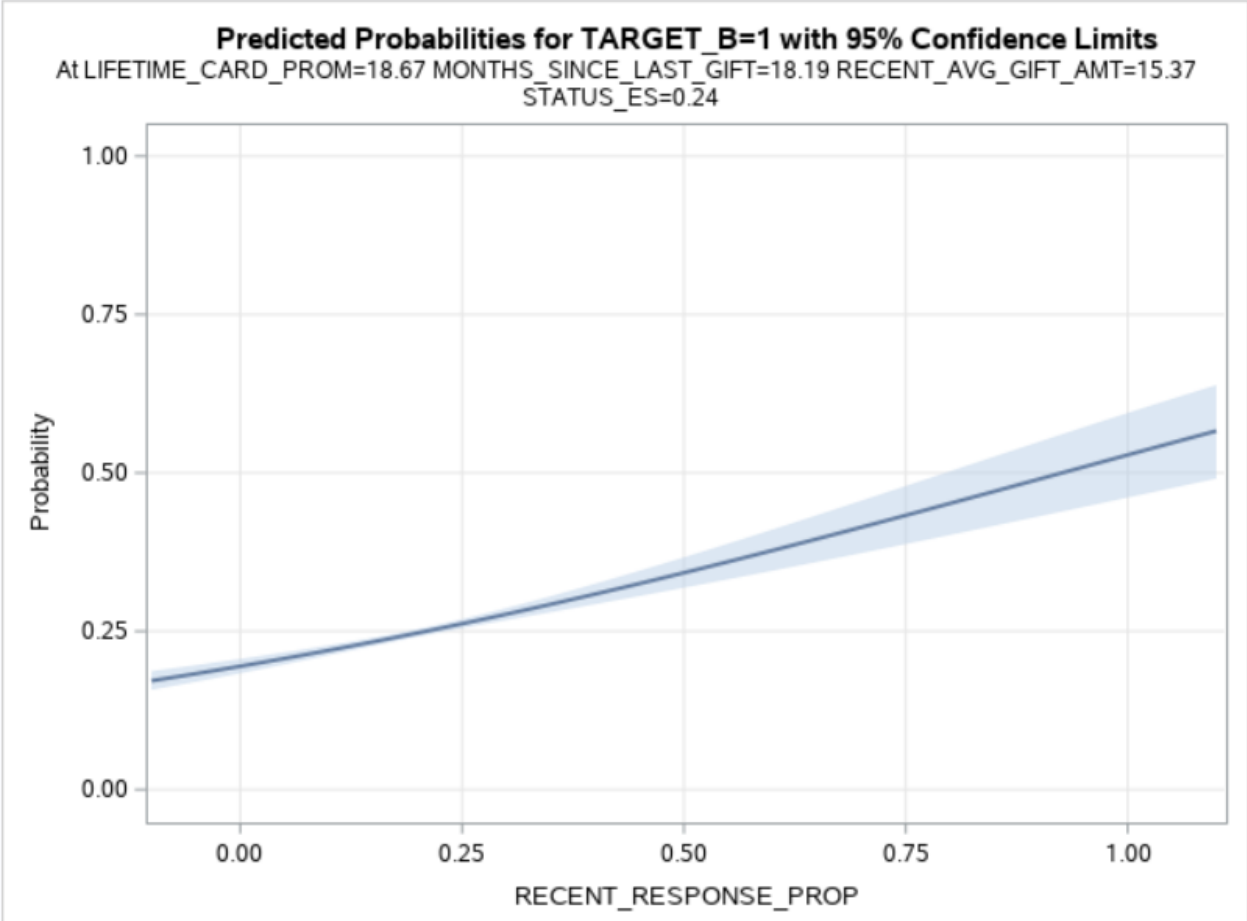
Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.9669	0.1131	73.0559	<.0001	
LIFETIME_CARD_PROM	1	0.0141	0.00216	42.8106	<.0001	0.0667
RECENT_RESPONSE_PROP	1	1.5328	0.1696	81.6972	<.0001	0.0963
MONTHS_SINCE_LAST_GIFT	1	-0.0325	0.00430	57.2378	<.0001	-0.0723
RECENT_AVG_GIFT_AMT	1	-0.0103	0.00206	25.1344	<.0001	-0.0580
STATUS_ES	1	0.1412	0.0453	9.7289	0.0018	0.0333

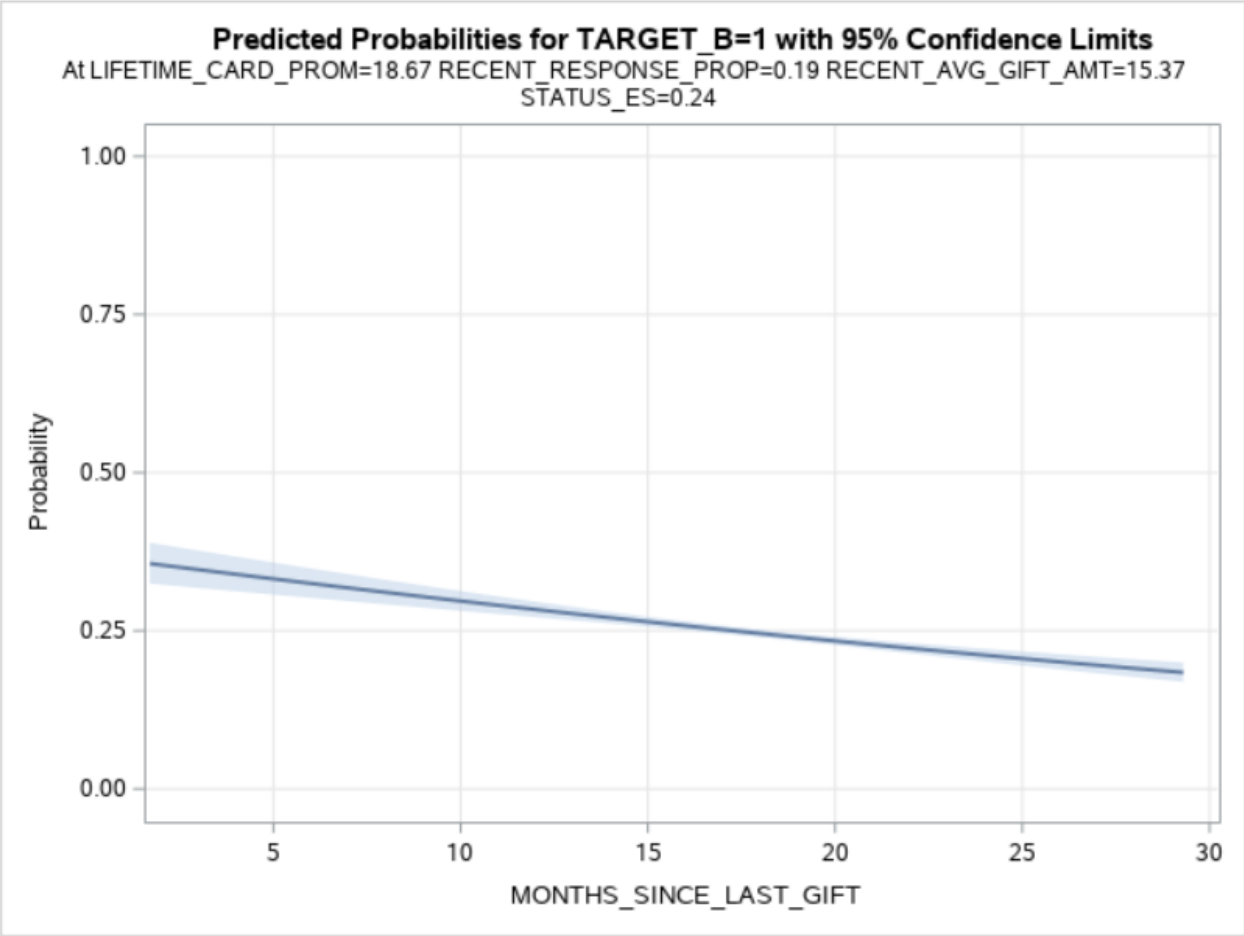
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	59.8	Somers' D	0.196
Percent Discordant	40.2	Gamma	0.196
Percent Tied	0.0	Tau-a	0.073
Pairs	70363947	c	0.598

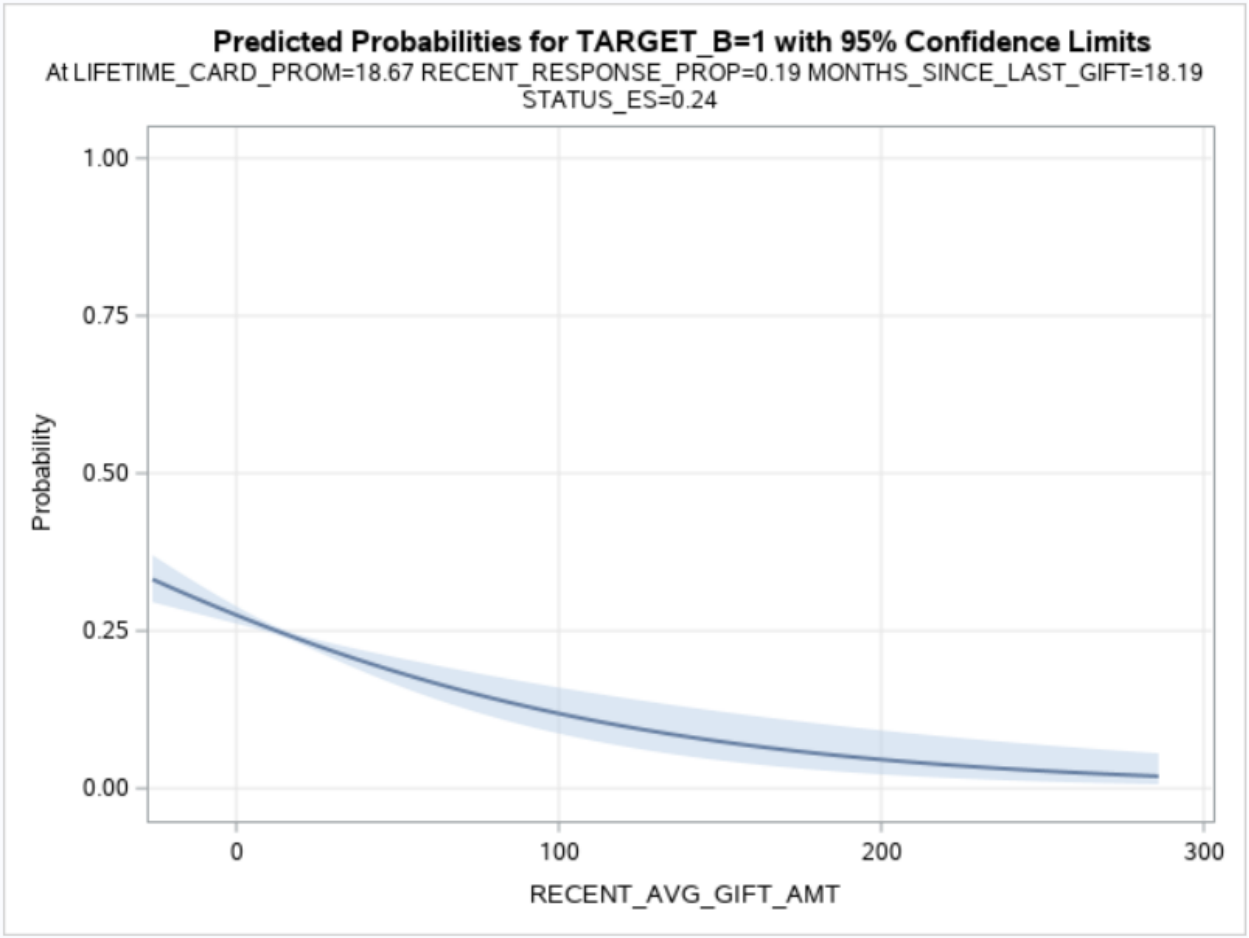
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
LIFETIME_CARD_PROM	10.0000	1.152	1.104	1.202
RECENT_RESPONSE_PROP	1.0000	4.631	3.321	6.457
MONTHS_SINCE_LAST_GIFT	6.0000	0.823	0.782	0.865
RECENT_AVG_GIFT_AMT	25.0000	0.772	0.697	0.853
STATUS_ES	1.0000	1.152	1.054	1.258

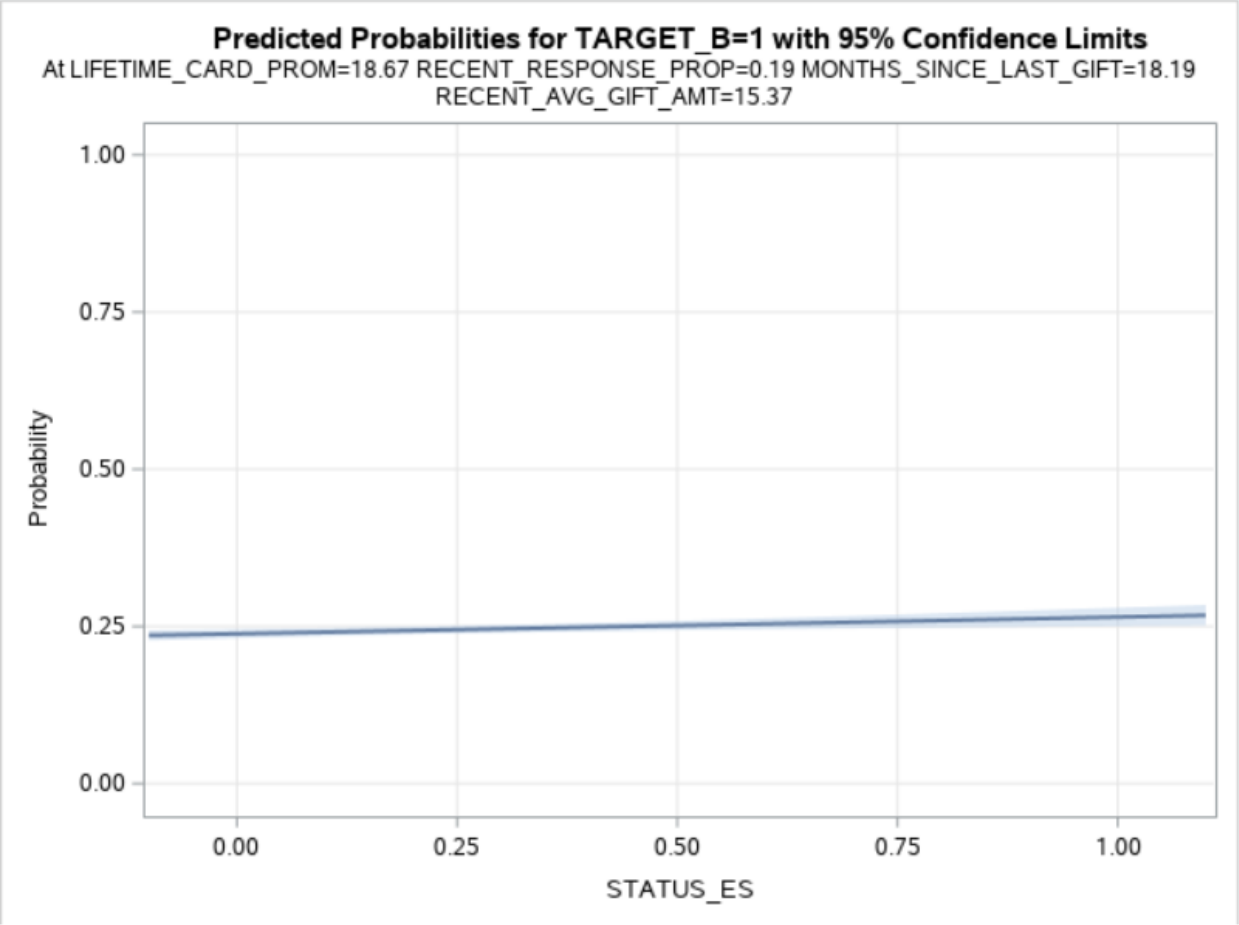

Odds Ratios with 95% Profile-Likelihood Confidence Limits
LIFETIME_CARD_PROM units=10
1.1519 (1.1041, 1.2018)
RECENT_RESPONSE_PROP
4.631 (3.3212, 6.457)
MONTHS_SINCE_LAST_GIFT units=6
0.8227 (0.7822, 0.8654)
RECENT_AVG_GIFT_AMT units=25
0.7721 (0.6966, 0.8527)
STATUS_ES
1.1517 (1.0537, 1.2584)
2
4
6
Odds Ratio


Predicted Probabilities for TARGET_B=1 with 95% Confidence Limits
At RECENT_RESPONSE_PROP=0.19 MONTHS_SINCE_LAST_GIFT=18.19 RECENT_AVG_GIFT_AMT=15.37 STATUS_ES=0.24
Probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
LIFETIME_CARD_PROM


Predicted Probabilities for TARGET_B=1 with 95% Confidence Limits
At LIFETIME_CARD_PROM=18.67 MONTHS_SINCE_LAST_GIFT=18.19 RECENT_AVG_GIFT_AMT=15.37 STATUS_ES=0.24
Probability
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
RECENT_RESPONSE_PROP


Predicted Probabilities for TARGET_B=1 with 95% Confidence Limits
At LIFETIME_CARD_PROM=18.67 RECENT_RESPONSE_PROP=0.19 RECENT_AVG_GIFT_AMT=15.37
STATUS_ES=0.24
Probability
1.00
0.75
0.50
0.25
0.00
5
10
15
20
25
30
MONTHS_SINCE_LAST_GIFT


Predicted Probabilities for TARGET_B=1 with 95% Confidence Limits
At LIFETIME_CARD_PROM=18.67 RECENT_RESPONSE_PROP=0.19 MONTHS_SINCE_LAST_GIFT=18.19 STATUS_ES=0.24
Probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
RECENT_AVG_GIFT_AMT


Predicted Probabilities for TARGET_B=1 with 95% Confidence Limits
At LIFETIME_CARD_PROM=18.67 RECENT_RESPONSE_PROP=0.19 MONTHS_SINCE_LAST_GIFT=18.19
RECENT_AVG_GIFT_AMT=15.37
Probability
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
STATUS_ES