

## Tutorial #3: Analysis of a Qualitative Predictor Variable

### DemoData = 'ex3.sav'

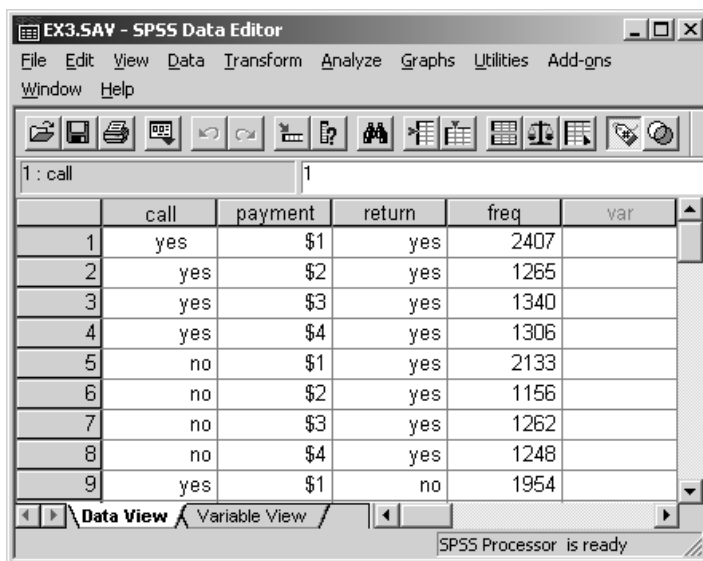
In Tutorial #2 *hypothetical* mail survey data were used to demonstrate the importance of assessing the model fit to test the validity of the assumptions made by a model. In the current tutorial we will analyze *real* data from the Mail Survey Experiment (Magidson, 1994b) where the actual dollar payments tested were \$1 (the control), \$2, \$3 and \$4, the experiment being designed so that there is no correlation between  $X_1$ :PAYMENT and  $X_2$ :CALL. We will again use the model fit statistic to choose between two alternative models, one that treats PAYMENT as a quantitative predictor with equidistant category scores (Model A) and one that treats PAYMENT as a qualitative predictor (Model B), which estimates the relative category spacing.

This tutorial illustrates

- How to assess the overall model fit
- How to determine whether fixed or free scores are best for a given categorical predictor
- The relationship between probabilities, odds and logits
- How to use the effects plot to interpret the beta effect estimates of the predictors under different contrast coding
- How to change the contrasts without re-estimating the model
- Use of the effects plot to compare the observed with the expected log-odds ratios

### The Data

The data for this file is in the SPSS file "ex3.sav". Figure 1 shows the first 9 cases:



	call	payment	return	freq	var
1	yes	\$1	yes	2407	
2	yes	\$2	yes	1265	
3	yes	\$3	yes	1340	
4	yes	\$4	yes	1306	
5	no	\$1	yes	2133	
6	no	\$2	yes	1156	
7	no	\$3	yes	1262	
8	no	\$4	yes	1248	
9	yes	\$1	no	1954	

Figure 1. Data file

The data from the Mail Survey Experiment is summarized in Table 1. The observed rate of returning the survey is given for each joint category formed by the joint predictor  $X$ =CALL x PAYMENT,

in percentages, odds and logit units. For example, among persons who received both a \$1 payment and a reminder call, 55.2% responded. The observed odds in favor of response for this group is  $55.2\% / (100\% - 55.2\%) = 1.23$  and the logit =  $\ln(1.23) = 0.21$ .

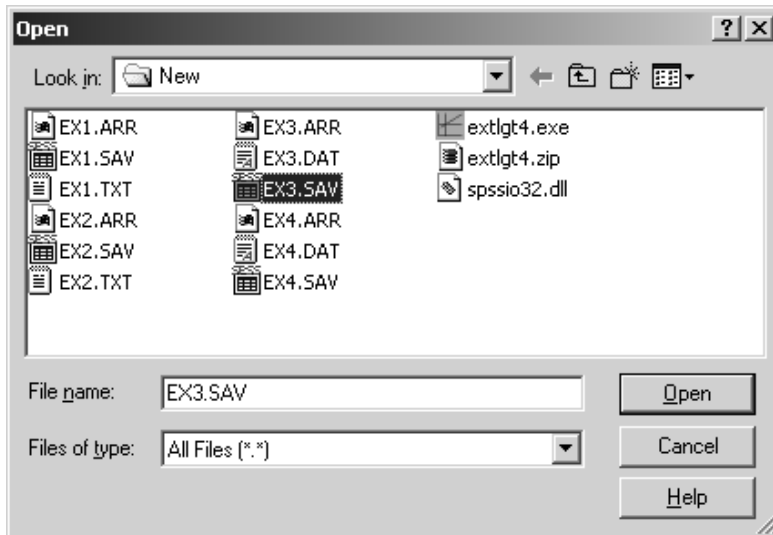
CALL	PAYMENT	RETURNED		Likelihood of Response		
		Yes	No	Percentage	Odds	Logit
Yes (1)	\$1 (1)	2,407	1,954	55.2%	1.23	0.21
	\$2 (2)	1,265	881	58.9%	1.44	0.36
	\$3 (3)	1,340	809	62.4%	1.66	0.5
	\$4 (4)	1,306	779	62.6%	1.68	0.52
No (0)	\$1 (1)	2,133	2,176	49.5%	0.98	-0.02
	\$2 (2)	1,156	942	55.1%	1.23	0.2
	\$3 (3)	1,262	897	58.5%	1.41	0.34
	\$4 (4)	1,248	839	59.8%	1.49	0.4

**Table 1. Mail Survey Experiment with Descriptive Statistics**

## The Model

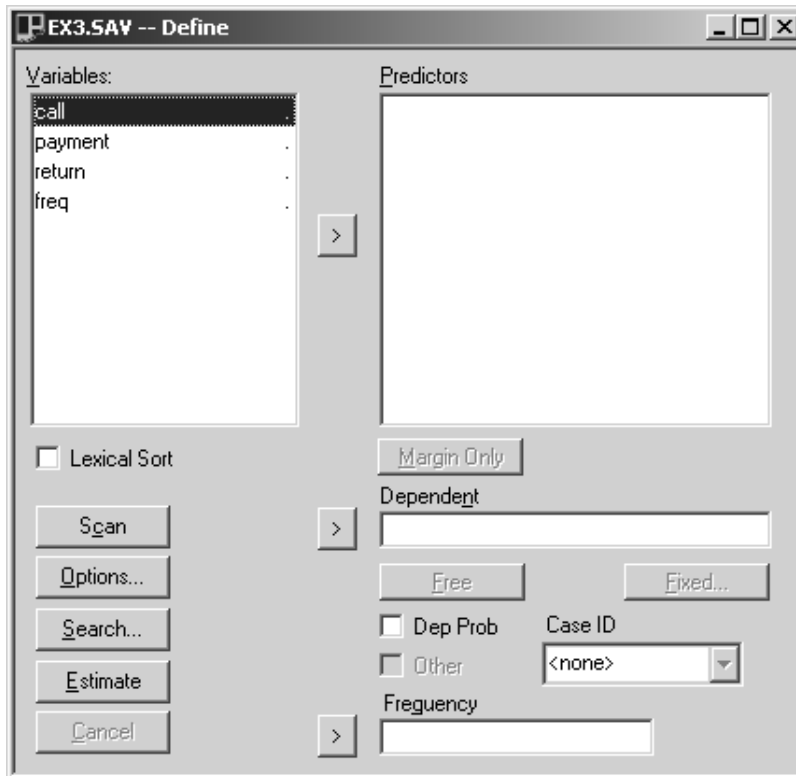
To open the data file in GOLDMineR,

- Click File, and then Open
- Pick ex3.sav and click OK



**Figure 2. File Open Dialog Box**

The Define Dialog Box opens:



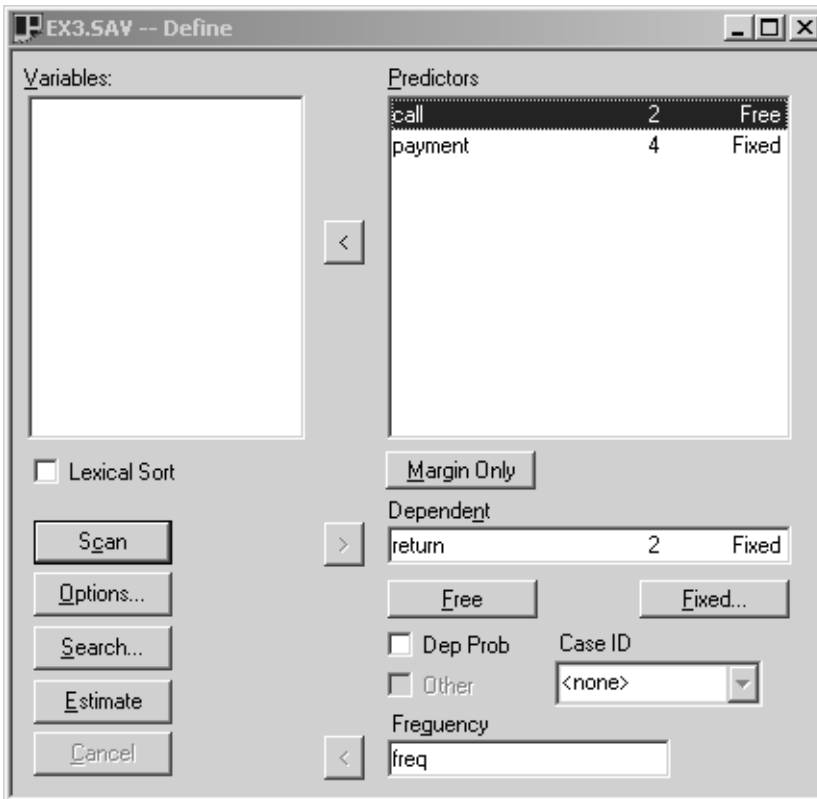
**Figure 3. The Define Dialog Box**

- Click on RETURN and press the arrow next to the Dependent box to move RETURN to the Dependent Variables box.
- Since CALL and PAYMENT are the predictor variables, select them and press the arrow next to the Predictors box to move them to the Predictors box.
- Click on FREQ and press the arrow next to the Frequency box to move FREQ to the Frequency box.
- Click Scan to scan the variables

By default, CALL is treated as a quantitative variable, using the dummy codes 1,0 as specified in the data file. While dichotomous variables may be treated as either 'Fixed' or 'Free' without altering the model statistics or model predictions, for improved clarity in the output, we will change the scale to 'Free' in which case both categories will appear in the statistics output by default.

- Right-click on CALL and select "Free" to convert CALL to a nominal variable.

The Define Dialog Box should now look like this:



**Figure 4. Define Dialog Box after Scan**

- Double-click on each of the predictors to examine the scores and labels associated with each of their categories.

We see that CALL has two levels – corresponding to YES and NO and PAYMENT four levels corresponding to \$1, \$2, \$3, and \$4:

Level	Label	Score	Count
1	no	0.000	10653
2	yes	1.000	10741

Level	Label	Score	Count
1	\$1	1.000	8670
2	\$2	2.000	4244
3	\$3	3.000	4308
4	\$4	4.000	4172

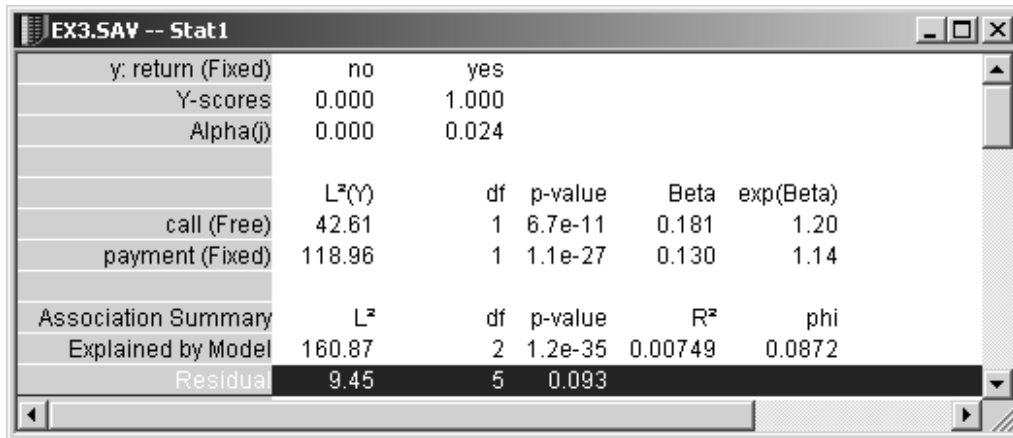
By default, PAYMENT is treated as a quantitative variable with the fixed equidistant category scores 1, 2, 3 and 4 as coded on the data file. This is indicated by 'Fixed' to the right of the variable name in the Define Dialog Box.

We will first estimate Model A, which treats PAYMENT as a quantitative predictor with the current equidistant category scores.

- Click Estimate to estimate the model

## Interpretation of Model Results

Here are the model summary results for Model A



The screenshot shows the 'Model Summary' window in SPSS. The window title is 'EX3.SAV -- Stat1'. It displays the following data:

	no	yes			
y: return (Fixed)	no	yes			
Y-scores	0.000	1.000			
Alpha(j)	0.000	0.024			
	L <sup>2</sup> (Y)	df	p-value	Beta	exp(Beta)
call (Free)	42.61	1	6.7e-11	0.181	1.20
payment (Fixed)	118.96	1	1.1e-27	0.130	1.14
Association Summary	L <sup>2</sup>	df	p-value	R <sup>2</sup>	phi
Explained by Model	160.87	2	1.2e-35	0.00749	0.0872
Residual	9.45	5	0.093		

**Figure 5. Model Summary Results of Model A which Specifies Equidistant Payment Scores**

Notice that overall, this model provides an adequate fit to the data. The Residual  $L^2 = 9.45$  (with 5 *df*), so that the p-value = .09.

Notice that the effect estimate for PAYMENT is beta = 0.130. To help interpret this quantity, denoted below as  $\hat{\beta}_1$ , we will now examine the partial effects plot for PAYMENT.

A partial regression plot is open by default.

- Click on the plot to make it the active window.

To change the default plot to the partial effects Plot associated with PAYMENT,

- In the Plot menu, select Partial X:

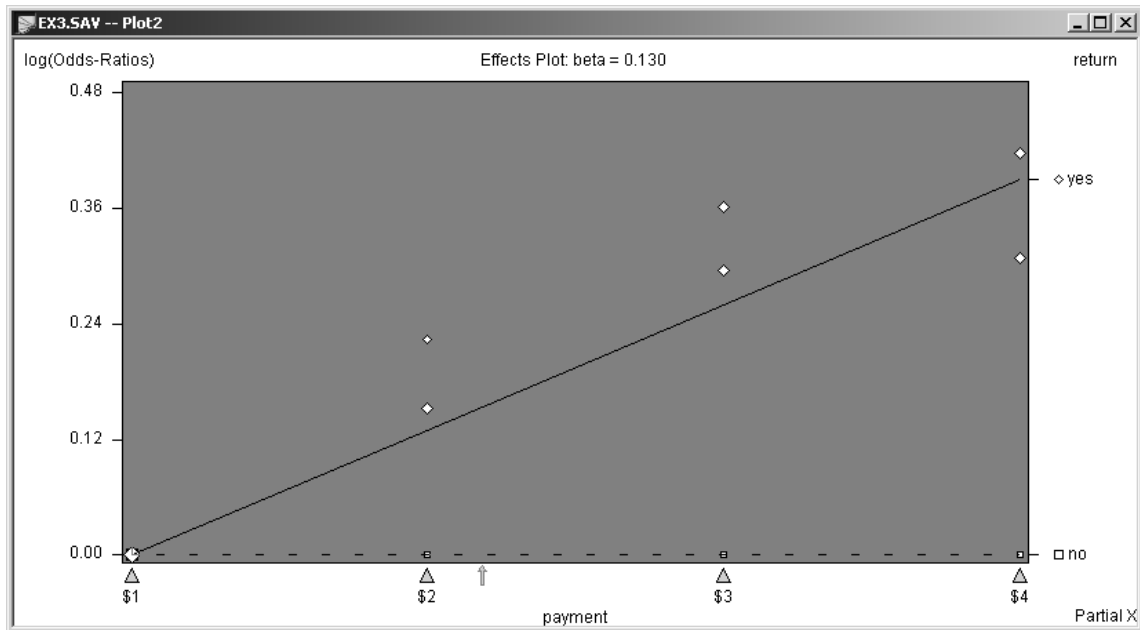


The Select Effect dialog box appears:



**Figure 6. Select Effect box**

- Double-click on Payment in the Select Effect box
- The partial regression plot as a function of PAYMENT appears:



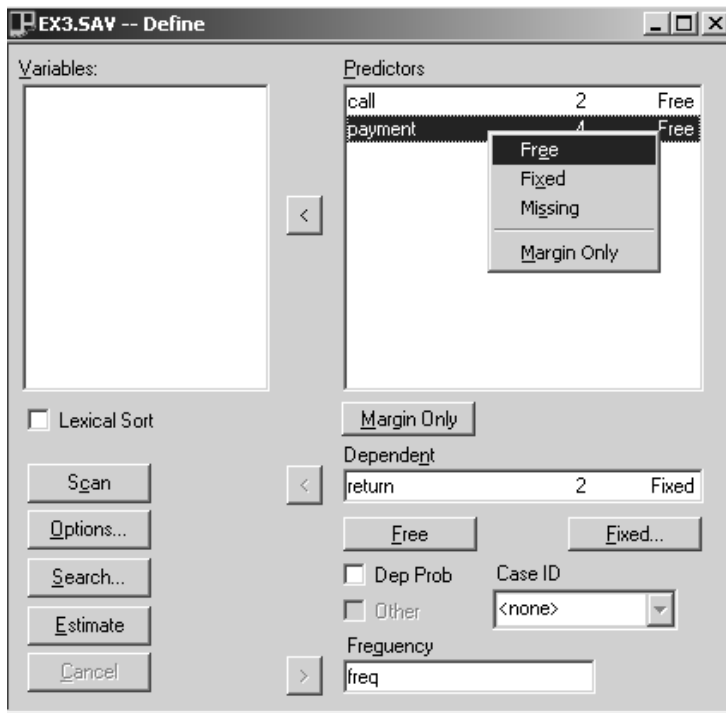
**Figure 7. Model A: PAYMENT is treated as Fixed with \$1 as the reference**

Figure 7 shows that persons receiving a \$2 payment and for whom CALL = Yes(No) are estimated to be  $\exp(0.33 \hat{\beta}_1) = 1.14$  times as likely to return the survey than persons receiving a \$1 payment and for whom CALL = Yes(No). Similarly, given CALL = Yes(No), the effect of a \$3 payment is estimated to be  $\exp(0.67 \hat{\beta}_1) = 1.30$  and a \$4 payment is  $\exp(\hat{\beta}_1) = 1.48$  times as likely to return the survey than persons receiving a \$1 payment (who were exposed to the same calling option).

Now, let us estimate Model B, which treats PAYMENT as a qualitative predictor. For that, we need to re-estimate the model using the Free scaling option for the PAYMENT variable. In order that the category scores estimated for any Free scale variable such as PAYMENT be identifiable, two standardizing restrictions must be imposed to uniquely define a zero point and a unit. By default, like Uniform scores, GOLDMineR restricts the scores by fixing the lowest at 0 and highest at 1.

To estimate Model B,

- In the Model menu, click Define
- Click on PAYMENT in the Predictors box
- Right-click and select Free from the pop-up menu



**Figure 8. Changing PAYMENT to Free**

- Click Estimate to estimate the Model

The overall model results for Model B are summarized in Figure 9

	L <sup>2</sup> (Y)	df	p-value	Beta	exp(Beta)
call (Free)	42.60	1	6.7e-11	0.181	1.20
payment (Free)	126.03	3	3.9e-27	0.363	1.44

Association Summary	L <sup>2</sup>	df	p-value	R <sup>2</sup>	phi
Explained by Model	167.94	4	2.9e-35	0.00783	0.0889
Residual	2.38	3	0.50		
Total	170.31	7	2.2e-33		

**Figure 9. Model B results**

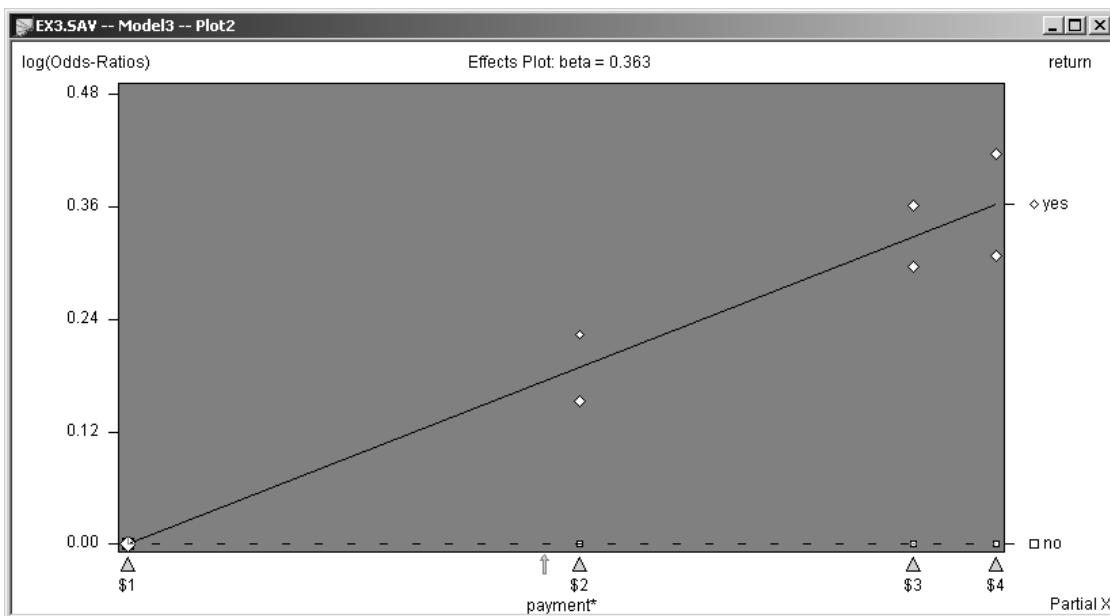
Scroll down to examine the Category-Specific Parameter Estimates:

**Figure 10. Category-Specific Parameter Estimates for Model B**

payment (Free)	X-score	C-weights	Beta(k)	Std. Err.	Wald	p-value	exp(Beta)	Lower	Upper
\$1	0.000	1.000	0.000	.	.	.	1.00	.	.
\$2	0.520	0.000	0.189	0.038	25.021	5.7e-7	1.21	1.12	1.30
\$3	0.906	0.000	0.329	0.038	75.335	4.0e-18	1.39	1.29	1.50
\$4	1.000	0.000	0.363	0.038	89.400	3.2e-21	1.44	1.33	1.55

Now, let us examine the Partial Effects plot for PAYMENT under Model B. The steps are similar:

- In the Plot menu, select Partial X
- Double-click on Payment in the Select Effect box
- The partial regression plot as a function of PAYMENT appears:



**Figure 11. Model B: PAYMENT is treated as a free predictor, with \$1 as the reference**


Figure 11 shows that persons receiving a \$2 payment and for whom CALL = Yes(No) are estimated to be  $\exp(0.52 \hat{\beta}_1) = 1.21$  times as likely to return the survey than persons receiving a \$1 payment and for whom CALL = Yes(No). Similarly, given CALL = Yes(No), the effect of a \$3 payment is estimated to be  $\exp(0.91 \hat{\beta}_1) = 1.39$  and a \$4 payment is  $\exp(\hat{\beta}_1) = 1.44$  times as likely to return the survey than persons receiving a \$1 payment (who were exposed to the same calling option).

Since PAYMENT is treated as a Free scale (qualitative) variable, GOLDMineR estimates category-specific effects  $\beta_k = \beta_{x_{1k}^*}$ , for each payment level  $k = 1, 2, 3, 4$ , and lists them in a separate section of the output. By default, dummy contrast coding is used with the category allocated the 0 score (the \$1 PAYMENT category) being the one omitted from the regression. The estimate for a \$2 payment, shown in Figure 10 as  $Beta(2) = 0.189$ , represents the increased likelihood of returning the survey associated with a \$2 payment vs. a \$1 payment. Examining the column titled p-value in Figure 10, we see that this effect is statistically significant ( $p=5.7 \times 10^{-7}$ ), and that  $\hat{\beta}_3$  and  $\hat{\beta}_4$  are also significantly different from zero.

To test the adequacy of the equidistant scoring assumption we compare the model fit statistics from the two models. Examining the output in Figure 5 and Figure 9 we see that Residual  $L^2$  has been reduced from 9.45 with 5 d.f. under Model A to 2.38 with 3 *df* under Model B, a reduction of 7.07 (with  $5-3 = 2$  *df*) which is statistically significant at the  $p = .05$  level. Hence, we reject the equidistant scaling assumption made by Model A in favor of Model B.

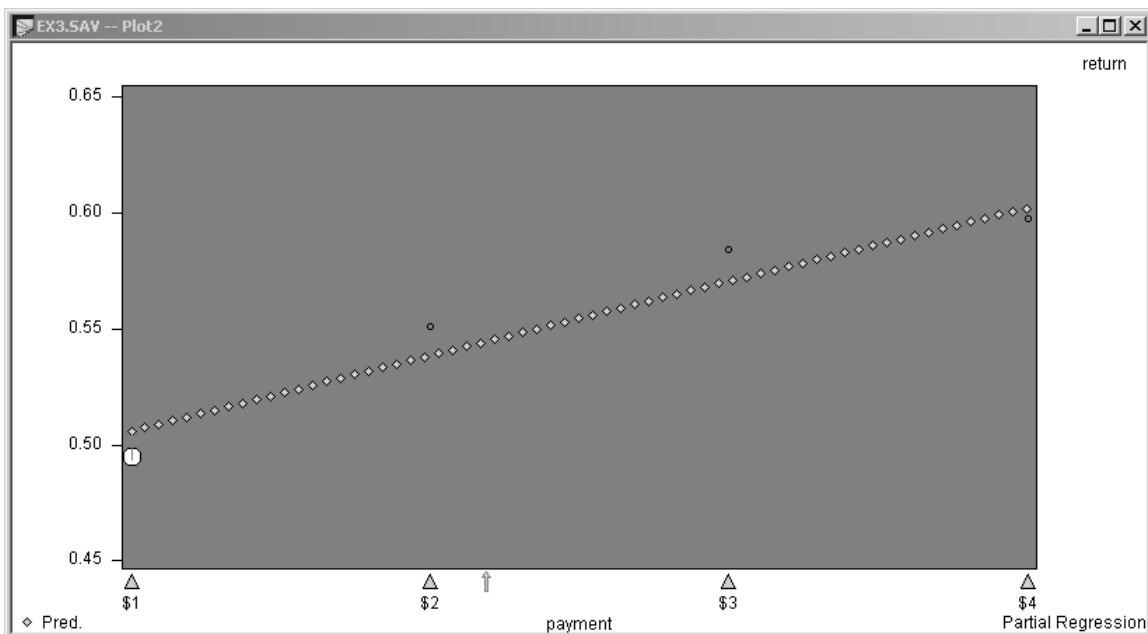
Suppose that we estimated only Model A. Would we still reject the equidistant scoring assumption? While Residual  $L^2 = 9.45$  (with 5 *df*) is an adequate fit ( $p = .09$ ), a comparison of the observed and expected log-odds ratio in Figure 7 might lead us to reexamine the equidistant scoring assumption. Specifically, the observed log-odds ratios for persons who received a reminder call, as well as the ratios for those who did not receive a call, both appear *above* the corresponding expected ratios (i.e., above the effects line) for each of the two middle payment categories (\$2 and \$3 payments). Note the improvement in fit attained in Figure 11 when the PAYMENT scores are no longer restricted to be equidistant.

To retrieve Model A

- In the Toolbar, click on the Summary icon: 
- Alternatively, you can select Summary Window from the Window menu
- The Summary Window appears:
- Double-click on Model1 (Model A)
- Select Plot from the pop-up menu

Now, let's view the plots:

- In the Plot menu, select Partial Regression
- Double-click on Payment in the Select Effect box

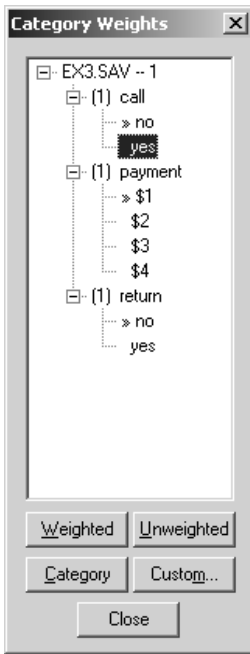


**Figure 12. Model A: Predicted Probability of RETURN = ‘Yes’ by PAYMENT given CALL = ‘No’**

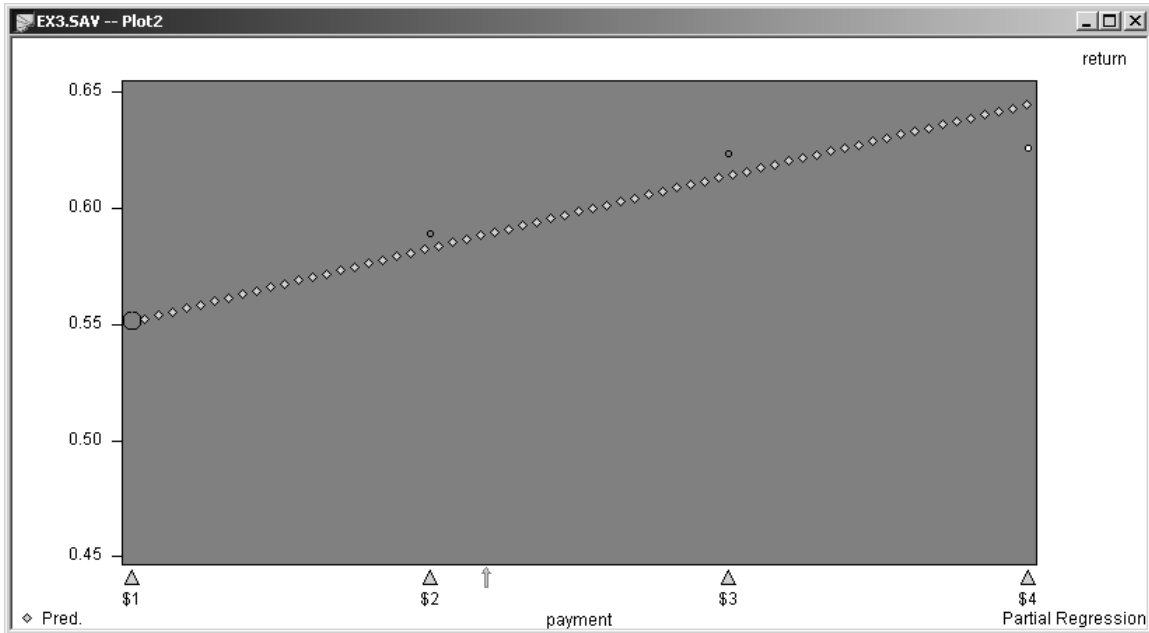
Now, let's switch CALL to “Yes” and observe the plot

- In the Model menu, select “Category Weights”

- The Category Weights box appears:



- Under CALL, double-click on Yes to change the contrast weights so that 'Yes' is the reference category
- The plot automatically changes:



**Figure 13. Model A: Predicted Probability of RETURN = 'Yes' by PAYMENT given CALL = 'Yes'**

Figures 12 and 13 illustrate partial Regression plots for PAYMENT given CALL under Model A. A careful examination of the pattern of residuals in these figures would also raise some questions about the correctness of the equidistant PAYMENT scores assumption. In particular, the two un-shaded circles indicate that the predicted return rates for the '\$1, no call' (see Figure 12) and '\$4, call' (see Figure 13)

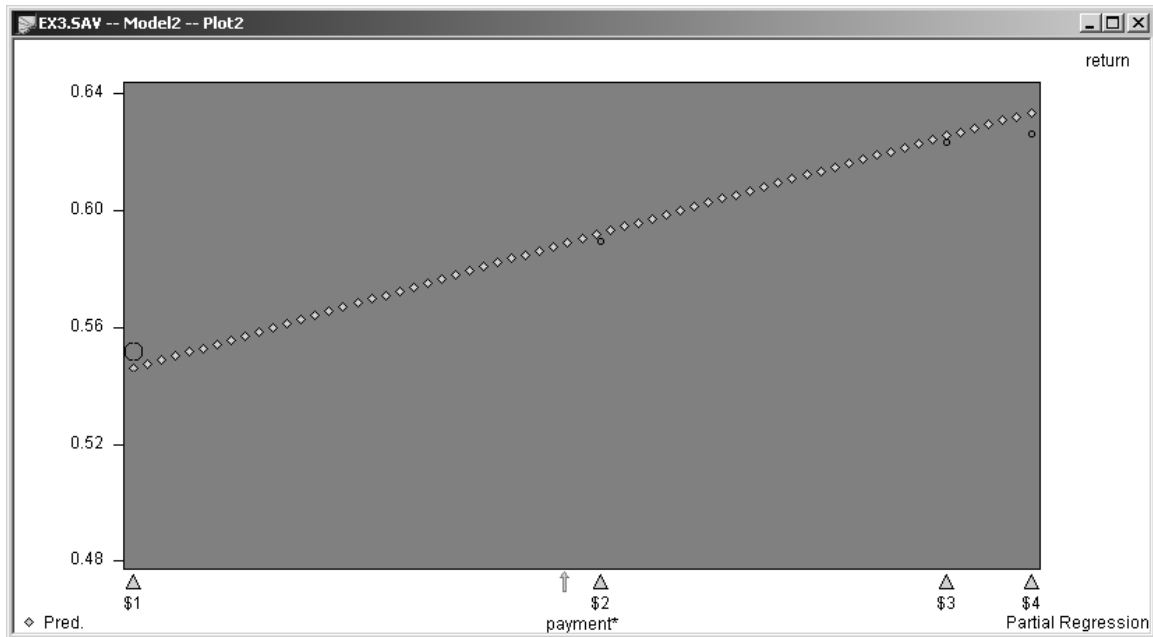
cells differ significantly from the observed rates. The other circles are shaded, indicating non-significant differences.

Now, let's go back to Model B:

- Click on the Summary icon to open the Summary Window
- Double-click on Model 2 .
- Select Plot from the pop-up menu

Let us observe the plots again:

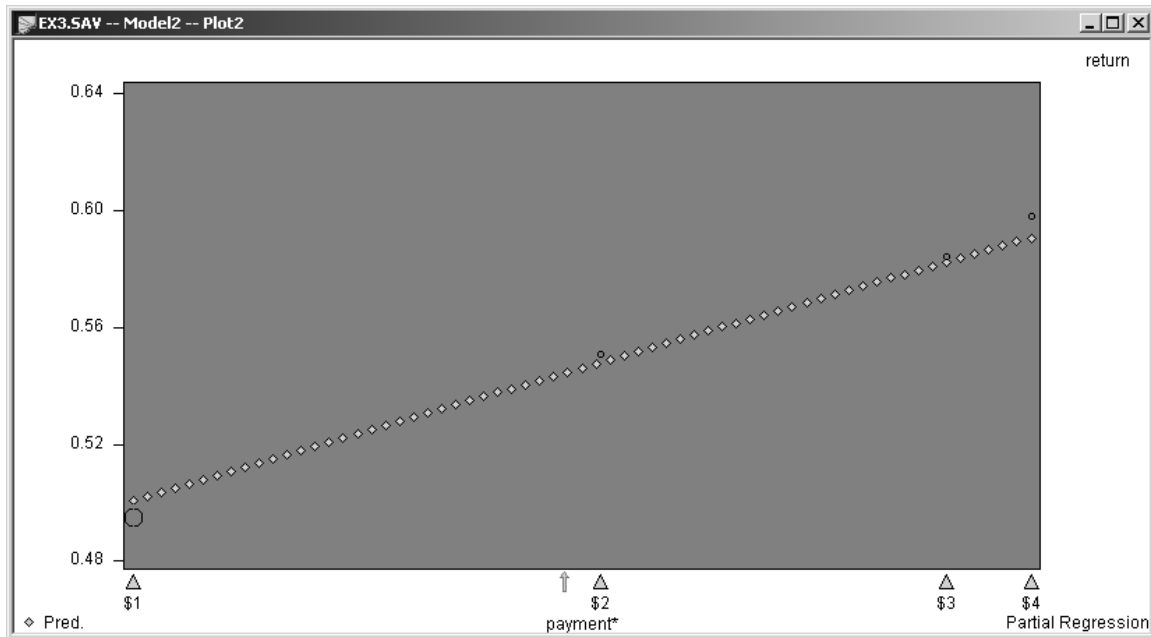
- In the Plot menu, select Partial Regression
- Double-click on Payment in the Select Effect box:



**Figure 14. Model B: Predicted Probability of RETURN = 'Yes' by PAYMENT given CALL = 'Yes'**

Note that the plot shows the probabilities associated with CALL= 'Yes', since that what was estimated previously. Now, once again,

- In the Model menu, select "Category Weights"
- Double-click on 'No' under CALL to switch to the CALL = 'No' situation:



**Figure 15. Model B: Predicted Probability of RETURN = ‘Yes’ by PAYMENT given CALL = ‘No’**

Figure 14 and Figure 15 show the closer fit to the observed return rates obtained under Model B where PAYMENT is treated as a qualitative variable (Free scale).

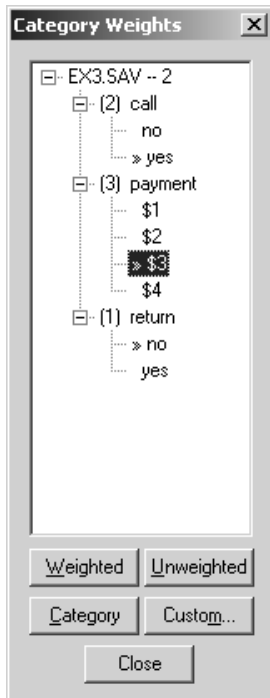
### Should we accept Model B as our final model?

The category-specific parameter estimate section of Figure 10 confirmed that a \$2 payment (as well as \$3 and \$4 payments) are significantly better than a \$1 payment ( $p = 5.7 \times 10^{-7}$ ). That is, all other factors being equal, persons receiving a \$2 payment are estimated to be 1.21 times as likely as those receiving a \$1 payment to return the survey. However, it is not clear from Figure 10 whether the estimate of  $\exp(\beta_4) = 1.44$  is significantly higher than the estimate of  $\exp(\beta_3) = 1.39$ . Similarly, from an examination of the Effects plot in Figure 11 we might ask the question of whether a \$4 payment is significantly better than a \$3 payment.

To test the effect of a \$4 payment relative to a \$3 payment we must change the reference category from the \$1 level to the \$3 level by the appropriate change in contrast coding. Once a model has been estimated, alternative contrast coding options may be selected to obtain different graphical views of the model and different category-specific parameter estimates, without re-estimating the model (i.e., only the category-specific parameter estimate section of the output changes).

To change the reference category,

- Click on the Model Summary window to make it active
- In the Model menu, select “Category Weights”
- Under PAYMENT, double-click on \$3 to make it the reference category:



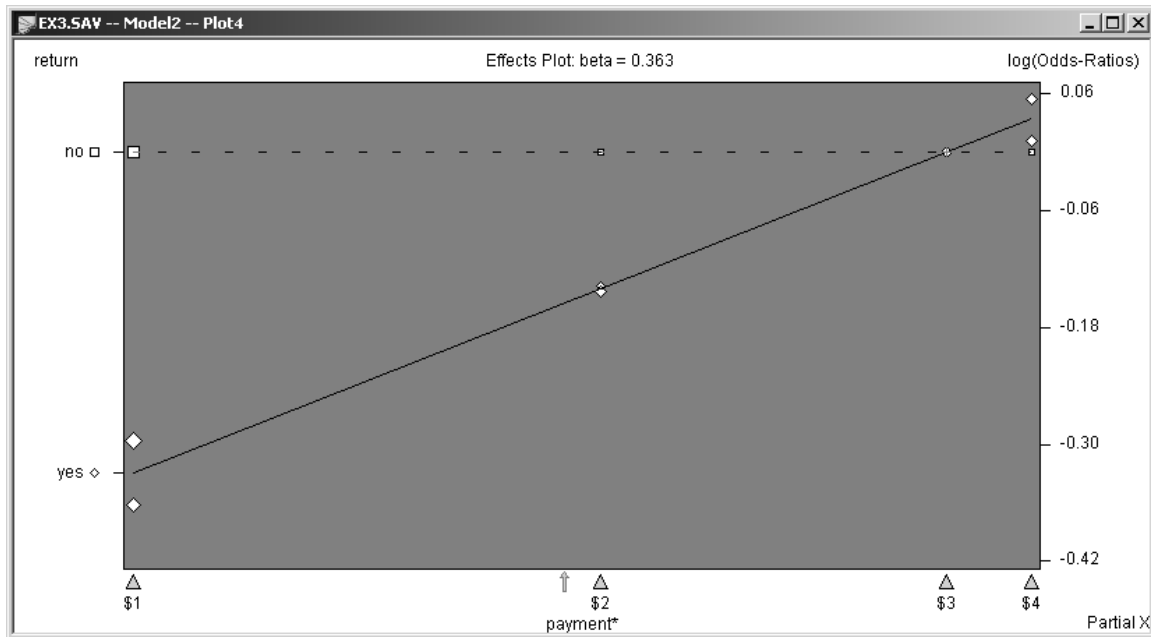
- Note that the numbers are automatically updated:

payment (Free)	X-score	C-weights	Beta(k)	Std. Err.	Wald	p-value	exp(Beta)	Lower	Upper
\$1	-0.906	0.000	-0.329	0.038	75.335	4.0e-18	0.72	0.67	0.78
\$2	-0.385	0.000	-0.140	0.044	10.115	0.0015	0.87	0.80	0.95
\$3	0.000	1.000	0.000	.	.	.	1.00	.	.
\$4	0.094	0.000	0.034	0.045	0.591	0.44	1.03	0.95	1.13

**Figure 16. Category-specific parameter estimates for PAYMENT under Model B where \$3 is the reference**

Now, let us observe the plot:

- In the Plot menu, select Partial X:



**Figure 17. Model B: PAYMENT is treated as a Free predictor with \$3 used as the reference to assess the significance of \$4 vs. \$3**

Figure 16 and Figure 17 provide results for Model B when the reference is changed to the \$3 level. Note that the X-scores in Figure 16 are lower than those in Figure 10 by 0.91, the value of the original X-score shown in Figure 10 for the \$3 payment level).

Figure 17 shows that persons receiving a \$4 payment and for whom CALL = Yes(No) are estimated to be  $\exp(0.09 \hat{\beta}_1) = 1.04$  times as likely to return the survey than persons receiving a \$3 payment and for whom CALL = Yes(No).

Figure 16 shows that the effect of \$4 relative to \$3,  $\hat{\beta}_4 = .034$ , is not significant at the .05 level ( $p = 0.44$ ). This result suggests that we assign the same score to a \$3 and a \$4 payment and estimate a final model. One possibility for a final model is to recode the original datafile to combine payment levels \$3 and \$4 and estimate a model with PAYMENT treated as either Free or Fixed using equidistant scores. Recoding could also be implemented by using the Group feature in GOLDMineR which will automatically combine the \$3 and \$4 payment levels if 3 groups are specified.

For the remainder of this tutorial, we maintain separate categories for the \$3 and \$4 payment levels and assign Fixed PAYMENT scores of 0, 0.5, 1 and 1. That is, “Model C” positions the \$2 payment equidistant between \$1 and \$3, and positions the \$4 payment at the same place on the plot as the \$3 payment.

To assign these new scores:

- In the Model menu, select Define
- In the Model Define Dialog Box, select PAYMENT
- Right-click and select Fixed from the pop-up menu
- Double-click on PAYMENT to open the category scores box
- Select the \$1 level
- In the Replace Box in the bottom left-hand corner, write in “0” and Click Replace:

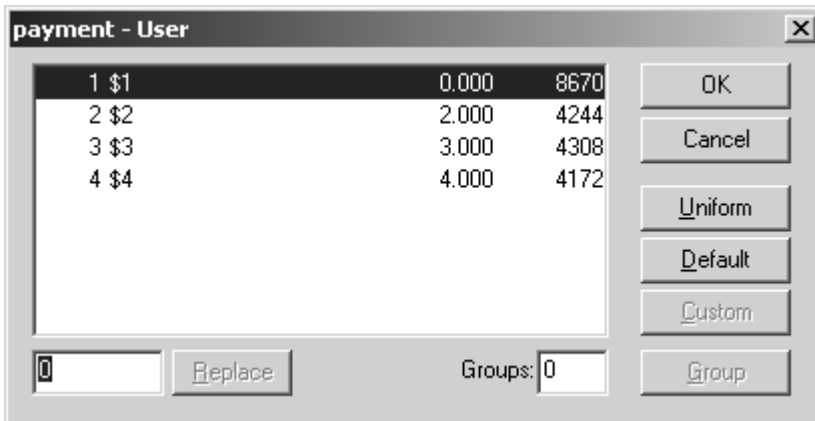


Figure 18. Assigning a score of “0” to level 1 of PAYMENT

- Repeat the procedure for levels 2-4, assigning “0.5” to level 2, “1” to level 3, and “1” to level 4

The final category scores box for PAYMENT should look like this:

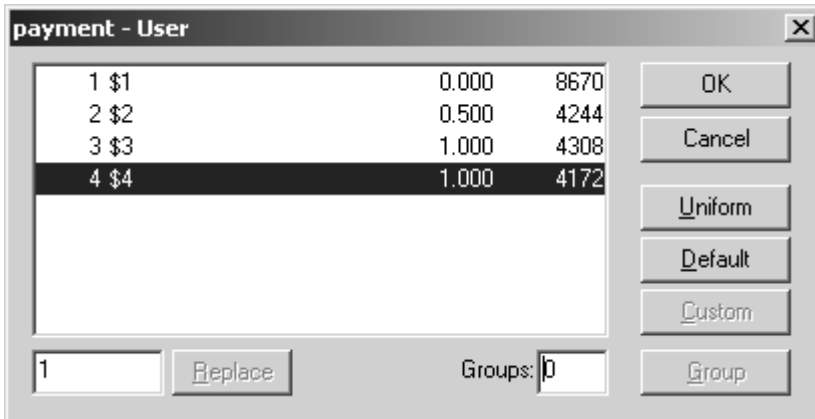


Figure 19. Category scores box for PAYMENT with new scores

- Click OK
- Click Estimate to re-estimate the Model:

Variable	L <sup>2</sup> (Y)	df	p-value	Beta	exp(Beta)
call (Free)	42.63	1	6.6e-11	0.181	1.20
payment (Fixed)	125.23	1	4.5e-29	0.346	1.41

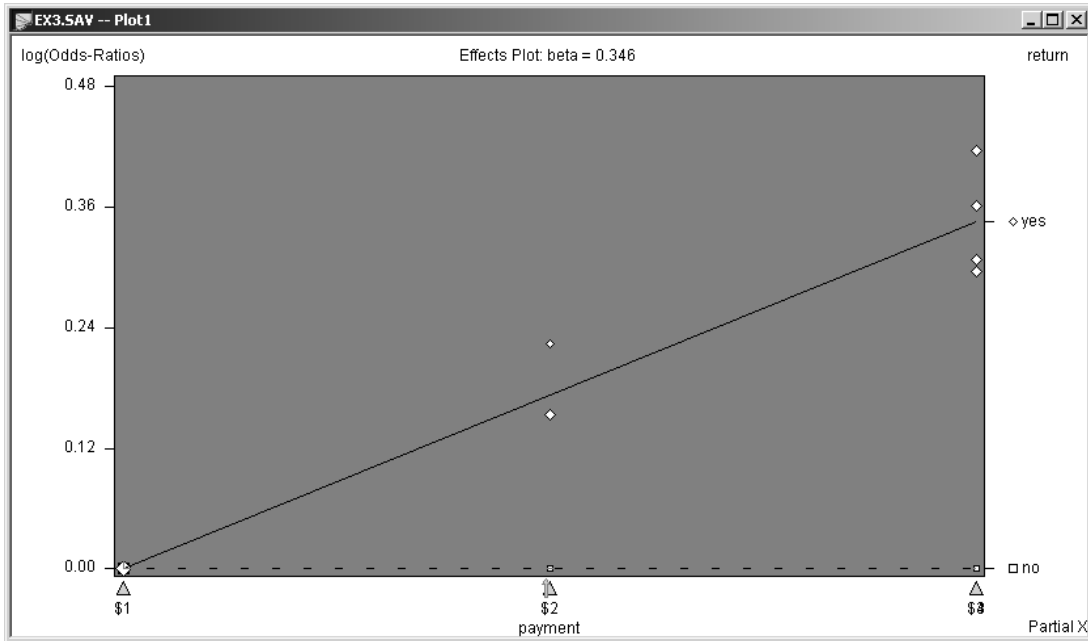
  

Association Summary	L <sup>2</sup>	df	p-value	R <sup>2</sup>	phi
Explained by Model	167.13	2	5.1e-37	0.00779	0.0887

Figure 20. Results of model specifying Fixed PAYMENT scores: 0, 0.5, 1, 1

Observe the plot:

- In the Plot menu, select Partial X
- In the Select Effects box, double-click on PAYMENT:



**Figure 21. Effects Plot for Model C specifying Fixed PAYMENT scores: 0, 0.5, 1, 1**

Figure 21 shows that persons receiving a \$3 or \$4 payment and for whom CALL = Yes(No) are estimated to be  $\exp(\hat{\beta}_1) = 1.41$  times as likely to return the survey than persons receiving a \$1 payment and for whom CALL = Yes(No).

Under Model C, each additional \$1 PAYMENT up to \$3 is expected to increase the likelihood of RETURN. A payment of \$4 is not expected to lift response above that of a \$3 payment.