Mert Dogan Assignment 6
Z1910844 10/23/2023

1. Use the dataset titled “FinalQ2” posted to solve this question. The dataset contains data on project proposal funding requests. Note that resource length summary pertains to length of the funding request in words, and TotalQuantity refers to the number of different line items requested for the project. Identify what factors would predict Project Approval (VAR7 0 = not approved, 1 = approved)

We are observing the factors that are predicting VAR7, which is whether a project is approved or not and listed as a binary categorical variable with values 0 or 1. Numerical independent variables to test are number of previously posted projects, resource summary length, total quantity, and total amount. We also have a categorical independent variable of grade category. Using these independent variables, we will conduct a logistic regression analysis to identify the factors predicting project approval.

From our model fit statistics table (Appendix A.1), we can see that the intercept only model -2LL is 42695.834 whereas the -2LL for the intercept and variables is 45765.019. To further determine whether the model is a good fit or not, we can look at the testing global null hypothesis table (Appendix A.2). We can observe that the p-value for the likelihood ratio, score, and Wald statistic are well below our nominal alpha of .05. Which indicates that we have a significant model.
We can also observe by looking at the R-square and max-rescaled R-square that the variables explain between 0.97% and 1.69% of the variation in project approval.

By checking the Type 3 Analysis of Effects table (Appendix A.3), we can determine which variables are significant predictors of project approval. We can observe that all the variables except the grade category are significant predictors of project approval with p-values less than alpha. We can conclude that all factors except the grade category predict project approval.

Next, we can interpret the significant coefficients on the Analysis of Maximum Likelihood Estimates table (Appendix A.4). We can observe that, while an increase in resource summary length and number of previously posted projects increases the odds of project approval, an increase in total quantity and total amount decreases the odds of project approval.

We can also interpret the Odds Ratio Estimates table (Appendix A.5) as when the number of previous projects increases by one, the resulting odds of project approval increase by (1.012 – 1) * 100 which is 1.2. So, there is an increase of 1.2 in the odds of project approval as the number of previous projects goes up by one. As the total quantity increases by one, the odds of project approval drop by -0.5, as the total amount increases by one, the odds of project approval drop by -0.012, and as the resource summary length increases by one, the odds of project approval increase by 0.016.

Finally, we can observe that the c value is 0.587 (Appendix A.6). Which means that our model classifies 58% of the rows in a correct way. Based on the independent variables, whether the project was approved or not matches the actual data 58% of the time. While not the best, we can conclude that the model is good.

2. Using the junkmail dataset in the SASHELP directory, examine if the three variables CapAvg, CapLong and CapTotal are useful in predicting whether an email is junk or not (Class variable: 0 = Not Junk and 1= Junk).

We are observing whether CapAvg, CapLong, and CapTotal are useful in predicting Class, which classifies an email as junk or not junk by using a logistic regression analysis.

From our model fit statistics table (Appendix B.1), we can see that the null model -2LL is 6170.153 whereas the -2LL for the model with predictors is 5154.327, far less than the null model, which gives us some cause to expect that our model is a good fit. To further determine whether the model is a good fit or not, we can look at the testing global null hypothesis table (Appendix B.2). We can observe that the p-value for the likelihood ratio, score, and Wald statistic are well below our nominal alpha of .05. Which indicates that we have a significant model.
We can also observe by looking at the R-square and max-rescaled R-square that the variables explain between 19.81% and 26.83% of the variation in project approval.

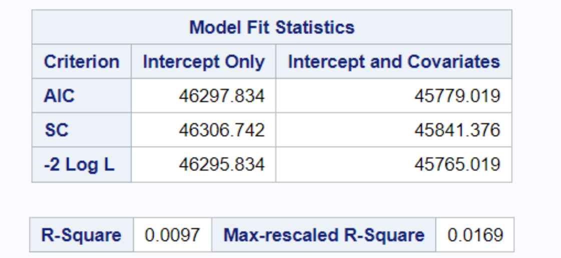
Next, we can interpret the coefficients on the Analysis of Maximum Likelihood Estimates table (Appendix B.3). We can observe that, while CapAvg and CapLong are good predictors of whether an email is junk or not, CapTotal is not a significant variable with a p-value greater than alpha. For CapAvg and CapLong, we can conclude that an increase in these variables increases the odds of the email being junk.

We can also interpret the Odds Ratio Estimates table (Appendix B.4) as when the number of CapAvg increases by one, the resulting odds of an email being junk increase by 11.3. Also, as the CapLong increases by one, the odds of an email being junk increase by 2.

Finally, we can observe that the c value is 0.806 (Appendix B.5). Which means that our model classifies 81% of the rows in a correct way. Based on the independent variables, whether the project was approved or not matches the actual data 81% of the time. We can conclude that this model is very good and CapAvg and CapLong are useful in predicting whether an email is junk or not.

APPENDIX A

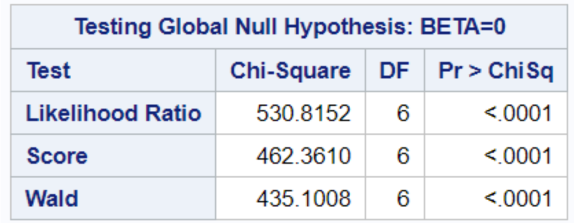
A.1



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	46297.834	45779.019
SC	46306.742	45841.376
-2 Log L	46295.834	45765.019

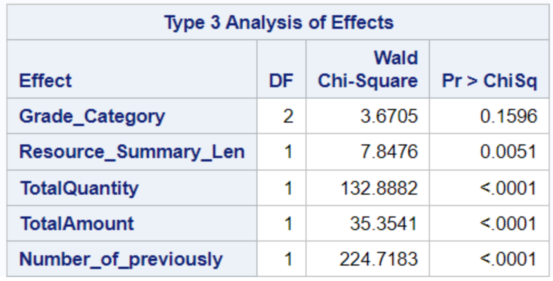
R-Square	0.0097	Max-rescaled R-Square	0.0169

A.2



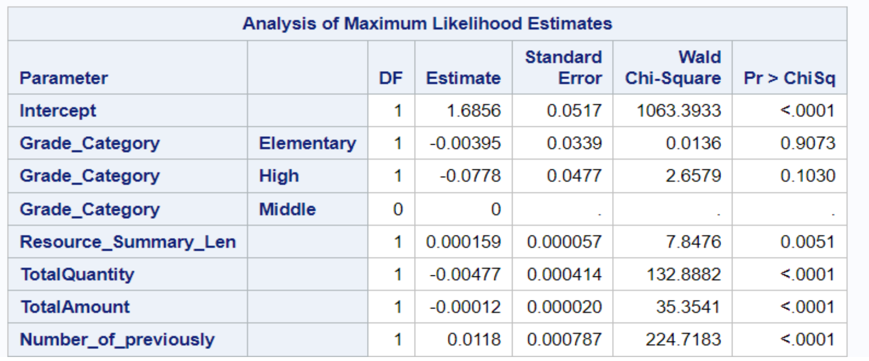
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	530.8152	6	<.0001
Score	462.3610	6	<.0001
Wald	435.1008	6	<.0001

A.3



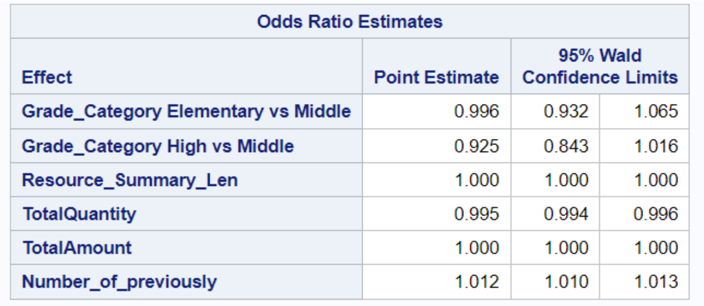
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Grade_Category	2	3.6705	0.1596
Resource_Summary_Len	1	7.8476	0.0051
TotalQuantity	1	132.8882	<.0001
TotalAmount	1	35.3541	<.0001
Number_of_previously	1	224.7183	<.0001

A.4



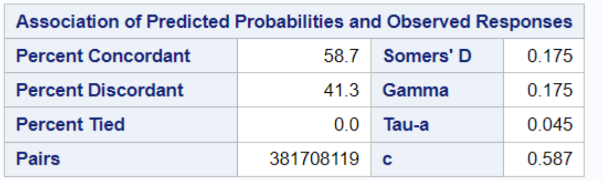
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.6856	0.0517	1063.3933	<.0001
Grade_Category	Elementary	1	-0.00395	0.0339	0.0136	0.9073
Grade_Category	High	1	-0.0778	0.0477	2.6579	0.1030
Grade_Category	Middle	0	0	.	.	.
Resource_Summary_Len		1	0.000159	0.000057	7.8476	0.0051
TotalQuantity		1	-0.00477	0.000414	132.8882	<.0001
TotalAmount		1	-0.00012	0.000020	35.3541	<.0001
Number_of_previously		1	0.0118	0.000787	224.7183	<.0001

A.5



Odds Ratio Estimates			
		95% Wald	
Effect	Point Estimate	Confidence Limits	
Grade_Category Elementary vs Middle	0.996	0.932	1.065
Grade_Category High vs Middle	0.925	0.843	1.016
Resource_Summary_Len	1.000	1.000	1.000
TotalQuantity	0.995	0.994	0.996
TotalAmount	1.000	1.000	1.000
Number_of_previously	1.012	1.010	1.013

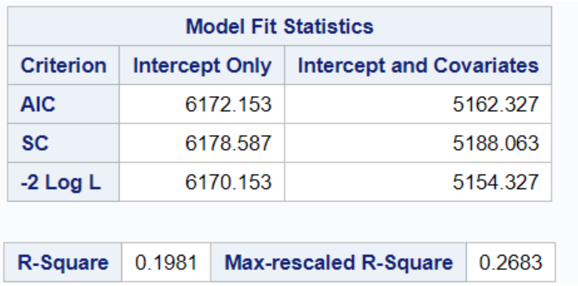
A.6



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	58.7	Somers' D	0.175
Percent Discordant	41.3	Gamma	0.175
Percent Tied	0.0	Tau-a	0.045
Pairs	381708119	c	0.587

APPENDIX B

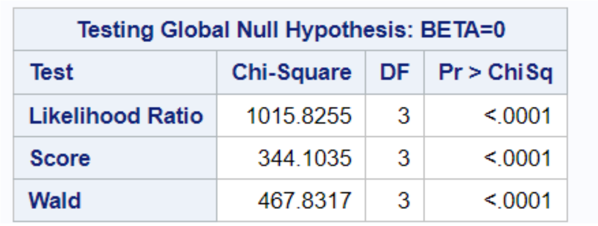
B.1



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6172.153	5162.327
SC	6178.587	5188.063
-2 Log L	6170.153	5154.327

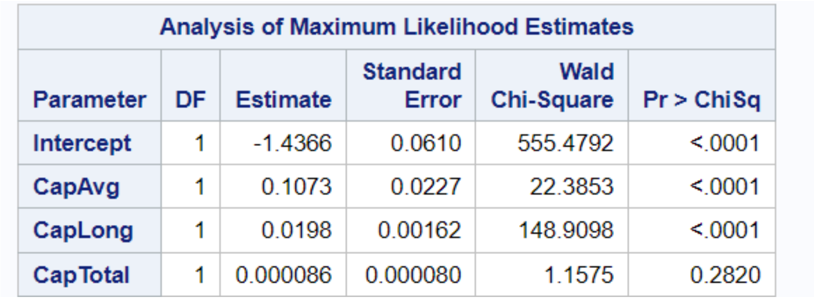
R-Square	0.1981	Max-rescaled R-Square	0.2683

B.2



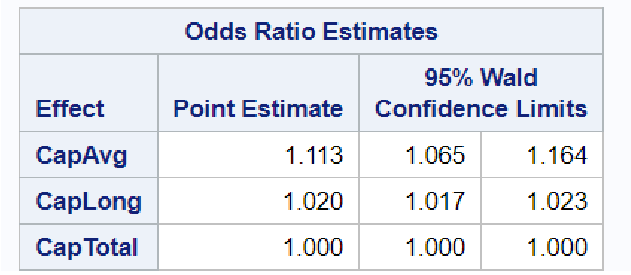
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1015.8255	3	<.0001
Score	344.1035	3	<.0001
Wald	467.8317	3	<.0001

B.3



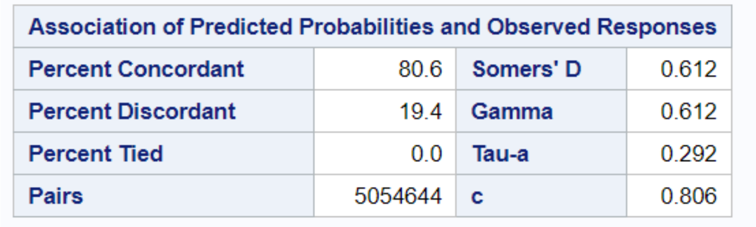
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4366	0.0610	555.4792	<.0001
CapAvg	1	0.1073	0.0227	22.3853	<.0001
CapLong	1	0.0198	0.00162	148.9098	<.0001
CapTotal	1	0.000086	0.000080	1.1575	0.2820

B.4



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
CapAvg	1.113	1.065	1.164
CapLong	1.020	1.017	1.023
CapTotal	1.000	1.000	1.000

B.5



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.6	Somers' D	0.612
Percent Discordant	19.4	Gamma	0.612
Percent Tied	0.0	Tau-a	0.292
Pairs	5054644	c	0.806