*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)*

1. The total number of observations read in the binary logistic regression model to determine which factors predict project approval is 54,624. The number of observations used is also 54,624. Of these observations, 8,227 projects were not approved for funding and 46,397 projects were approved. The factors analyzed include: grade category, number of previously posted projects, resource summary length (number of words in request), total quantity (total line items), and total amount. Grade category is our only categorical variable with three categories (elementary, high, middle).
2. The -2 log likelihood value for the intercept (null) is 46,295.834 which is greater than the -2 log likelihood value of the model with the factor predictors of 45,765.019. Also, the likelihood ratio, score, and wald tests each produce chi-square values and ultimately p-values of less than 0.0001 for each test. For all the reasons described above we can conclude this model to be a good fit. Although, the difference in -2 log likelihood is minimal. This aligns with the interpretation of the R-squared values. The model shows that the variation in project approval can be explained by between 0.97% and 1.69% of the predictor factors. These are low percentages.
3. When looking at the 3 analysis of effects table we can conclude which factors are significant predictors of project approval. The p-value for grade category is 0.1596 which is higher than an alpha value of 0.05 and therefore we conclude that grade category is not a significant predictor of project approval. The remaining p-values fall between less than 0.0001 and 0.0051 which is below 0.05. Therefore, number of previously posted projects, resource summary length, total quantity, and total amount are significant predictors of project approval.
4. The estimates for grade category are not interpreted based on the p-value. The estimates for number of previously posted projects and resource summary length are 0.0118 and 0.000159 meaning that as both of these factors increase so does their chance of project approval by the stated amount. On the other hand, total quantity and total amount have values of -0.00477 and 0.00012 meaning that as the total quantity and total quantity increase, the chance of project approval decreases by the stated amount.
5. The odds ratio estimates for each factor are as follows:
   1. Elementary vs. Middle: 0.996 -1 \*100 = -0.4
   2. High vs. Middle: 0.925 – 1 \* 100 = -7.5
   3. Number of Previously: 1.012 – 1 \* 100 = 1.2
   4. Resource Length Summary: 1.000 – 1 \* 100 = 0
   5. Total Quantity: 0.995 – 1 \* 100 = -0.5
   6. Total Amount: 1.000 – 1 \* 100 = 0
      1. As shown by the odds, there isn’t much movement as far as an increase in the odds of project approval based on the factors. Only number of previously posted projects displays a positive number of 1.2 and displays that as the number of previously posted projects increase by 1 the odds of project approval increase by 1.2. The remaining factors hold odds values of either 0 or negative numbers displaying a negative relationship in the odds as these factors increase. As an example, the odds of middle grade decrease by 7.5 times for every 1 increase in high grade.
6. Lastly, we interpret the c value. The c value is 58.7% which is greater than the minimal value of 50% and we can conclude the model is a good fit.

*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).*

1. The total number of observations read in the binary logistic regression model to determine which factors predict whether an email is junk or not is 4,601. The number of used observations is 1,813 which align with our selected variables. The variables include CapAvg, CapLong, and CapTotal.
2. The -2 log likelihood value for the intercept is 6170.153 and the -2 log likelihood value for the model with variables is 5154.327. The p-value for likelihood ratio, score, and wald tests are all less than 0.0001 which is less than an alpha value of 0.05. For both of these reasons we conclude the model to be a good fit. Also, according to our R-squared values, the variation in junk mail can be explained by 19.81% and 26.83% of the predictor variables. While these percentages are not very high, the model explains more of our dependent variable than the last problem.
3. When looking at the analysis of maximum likelihood estimates table we can conclude which factors are significant predictors of project approval. First, the intercept has a p-value of less than 0.0001 which is less than an alpha value or 0.05 which means the intercept can be interpreted. Also, CapAvg and CapLong have p-values less than 0.0001 which is also less than alpha and can be classifies as significant variables. However, CapTotal has a p-value of 0.2820 which is greater than alpha and therefore cannot be classified as a significant variable when predicting whether and email is junk or not.
4. CapAvg holds an estimate value of 0.1073 which is greater than CapLong with a value of 0.0198 meaning CapAvg has a greater impact on predicting junk mail as junk mail increases.
5. The odds ratio estimates are as follows:
   1. CapAvg: 1.113 – 1 \* 100 = 11.3
   2. CapLong: 1.020 – 1 \* 100 = 2
   3. CapTotal: 1.000 – 1 \* 100 = 0
   4. As CapAvg increases by 1 there the odds of receiving junk mail via email increase by 11.3 times. This is the greatest odds for junk mail of the three variables and it makes sense. Logically there should be more junk mail as the average number of mail increases. Similarly, as CapLong increases by 1, the odds of receiving junk mail via email increases by 2. The odds of CapTotal does not increase nor decrease as the odds are 0.
6. The c value holds a value of 80.6% and we can conclude that the model is a good fit. This is a much higher c value than the previous problem.

Appendix

Question 1:

| **Model Information** | | |
| --- | --- | --- |
| **Data Set** | \_TEMP0.FINALQ2 |  |
| **Response Variable** | VAR7 | Approved (0= Not approved, 1 = Approved) |
| **Number of Response Levels** | 2 |  |
| **Model** | binary logit |  |
| **Optimization Technique** | Fisher's scoring |  |

|  |  |
| --- | --- |
| **Number of Observations Read** | 54624 |
| **Number of Observations Used** | 54624 |

| **Response Profile** | | |
| --- | --- | --- |
| **Ordered Value** | **VAR7** | **Total Frequency** |
| **1** | 0 | 8227 |
| **2** | 1 | 46397 |

**Probability modeled is VAR7='1'.**

| **Class Level Information** | | | | |
| --- | --- | --- | --- | --- |
| **Class** | **Value** | **Design Variables** | | |
| **Grade\_Category** | **Elementary** | 1 | 0 | 0 |
|  | **High** | 0 | 1 | 0 |
|  | **Middle** | 0 | 0 | 1 |

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

| **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 |

| **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 |

| **Type 3 Analysis of Effects** | | | |
| --- | --- | --- | --- |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Grade\_Category** | 2 | 3.6705 | 0.1596 |
| **Number\_of\_previously** | 1 | 224.7183 | <.0001 |
| **Resource\_Summary\_Len** | 1 | 7.8476 | 0.0051 |
| **TotalQuantity** | 1 | 132.8882 | <.0001 |
| **TotalAmount** | 1 | 35.3541 | <.0001 |

| **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 | . | . | . |
| **Number\_of\_previously** |  | 1 | 0.0118 | 0.000787 | 224.7183 | <.0001 |
| **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 |

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **Grade\_Category Elementary vs Middle** | 0.996 | 0.932 | 1.065 |
| **Grade\_Category High vs Middle** | 0.925 | 0.843 | 1.016 |
| **Number\_of\_previously** | 1.012 | 1.010 | 1.013 |
| **Resource\_Summary\_Len** | 1.000 | 1.000 | 1.000 |
| **TotalQuantity** | 0.995 | 0.994 | 0.996 |
| **TotalAmount** | 1.000 | 1.000 | 1.000 |

| **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 |

Question 2:

| **Model Information** | | |
| --- | --- | --- |
| **Data Set** | SASHELP.JUNKMAIL | Classifying Email as Junk or Not |
| **Response Variable** | Class | 0 - Not Junk, 1 - Junk |
| **Number of Response Levels** | 2 |  |
| **Model** | binary logit |  |
| **Optimization Technique** | Fisher's scoring |  |

|  |  |
| --- | --- |
| **Number of Observations Read** | 4601 |
| **Number of Observations Used** | 4601 |

| **Response Profile** | | |
| --- | --- | --- |
| **Ordered Value** | **Class** | **Total Frequency** |
| **1** | 0 | 2788 |
| **2** | 1 | 1813 |

**Probability modeled is Class=1.**

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

| **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 |

| **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 |

| **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 |

| **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 |

| **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 |