Project 5: Logistic Regression with PUMS data

PUMS Data: Washington

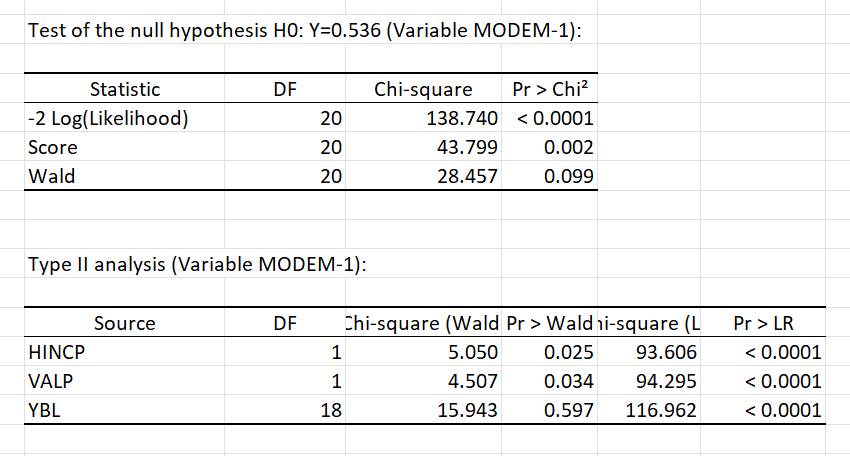
Katie Schreck, Cameron Dye, Qi Xue, Marina Dooley

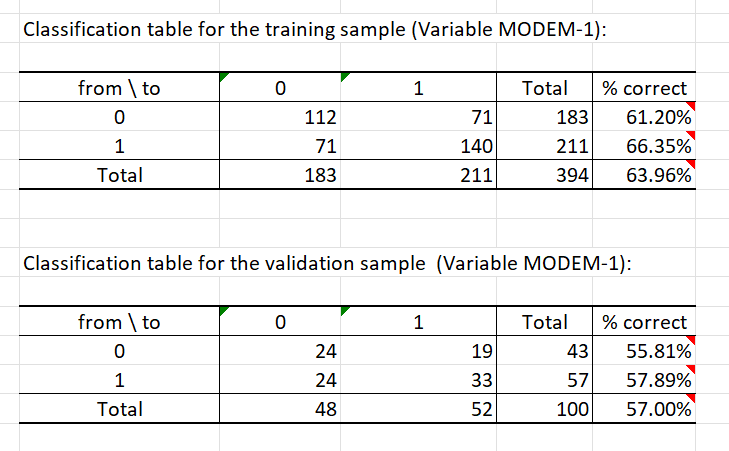
1. A brief explanation of your data (1-2 paragraphs):

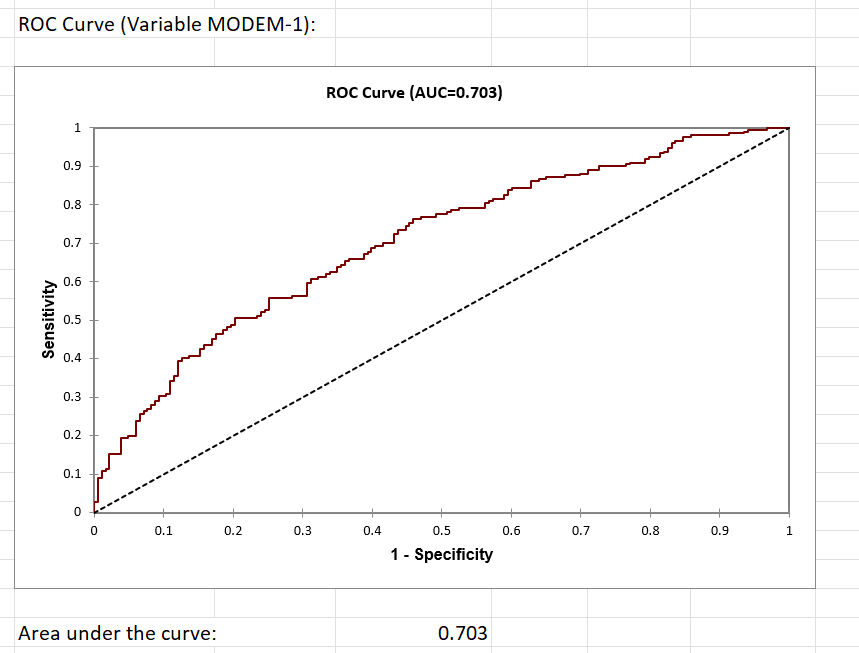
Our group used the data we selected for project number three which focuses on the state of Washington and we used four variables to predict the household income and now we will use the data to predict the outcome of a binary variable through logistic regression. The numerical variables we selected include the property value (VALP) and the response variable which is household income (HINCP). The categorical variables selected are family employment status (FES) as well as the year the property was built (YBL). These variables can help to predict household income in many ways. The categorical variable, family employment status, will allow us to understand if the family is employed and how many of them are bringing in money. We did some random sampling with our data and the total number of observations in our final data is 494.

1. Comparison of logistic regression models. Use a table similar to the one below:

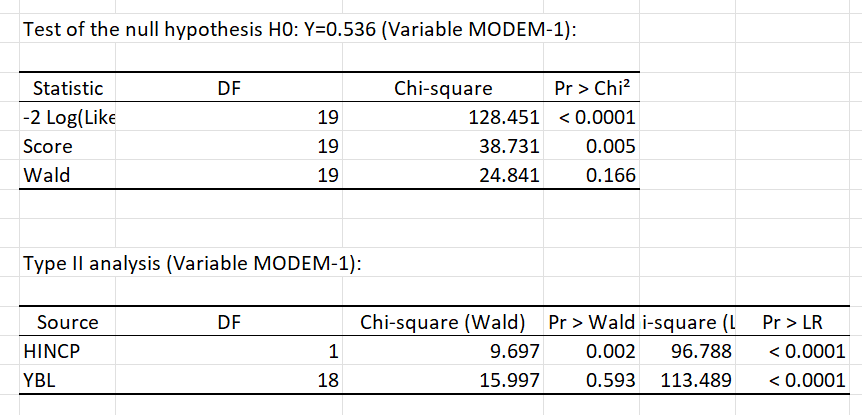
Model 1

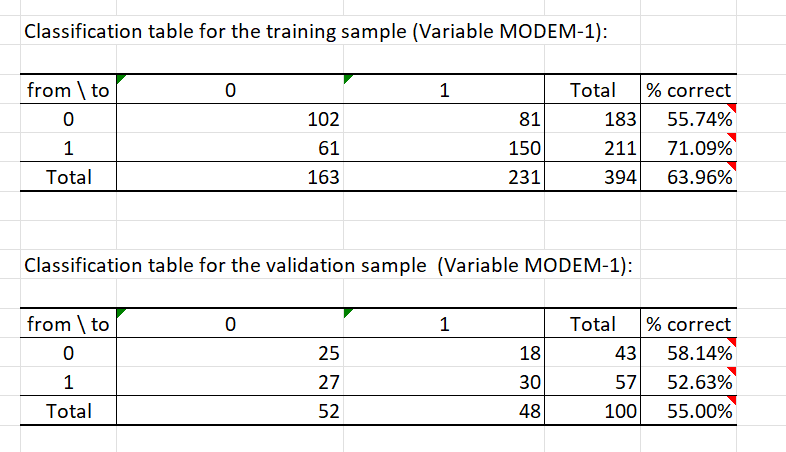


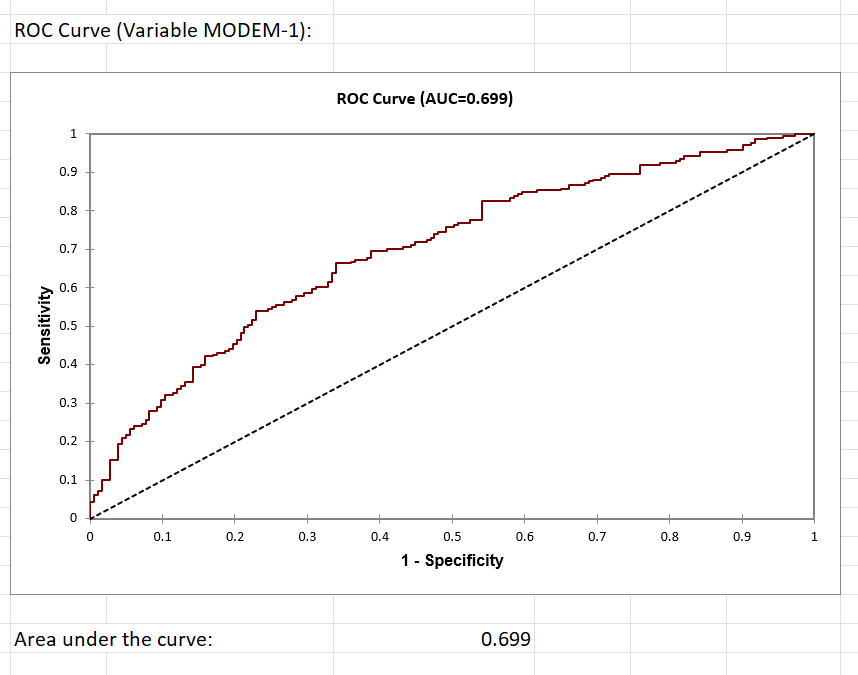




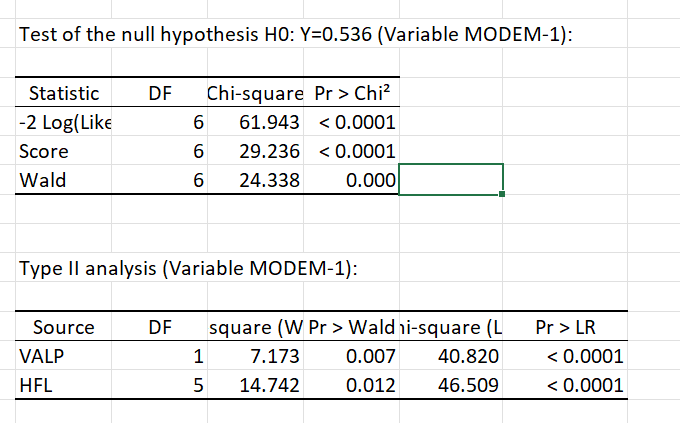
Model 2

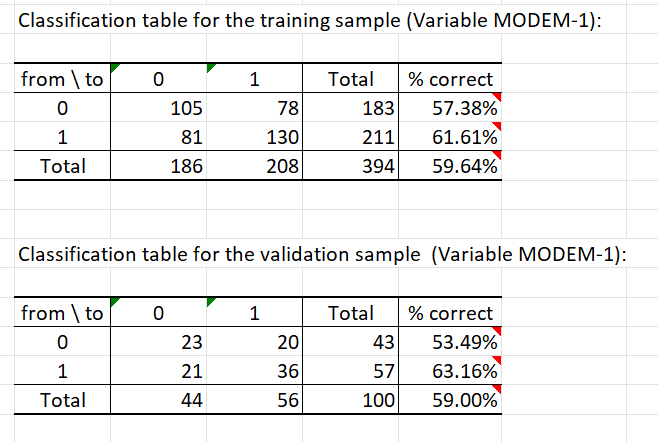


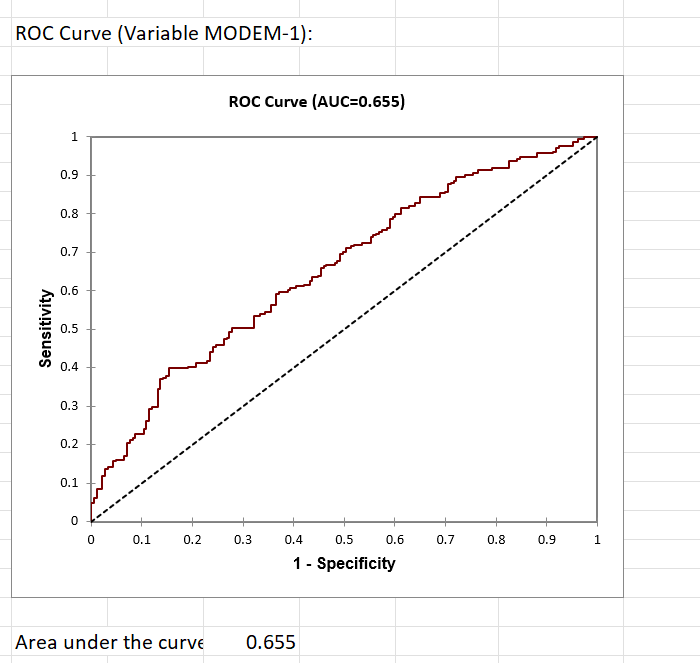




Model 3







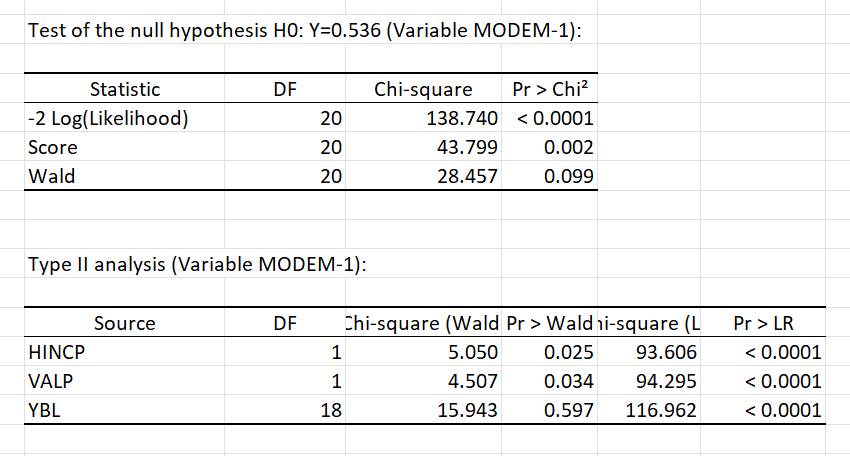
|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| Explanatory Variables Included | HINCP,VALP,YBL | HINCP,YBL | VALP,HFL |
| Are all the slope coefficients significant? (Hint: Compare *p*-values with 5%) | Yes | Yes | Yes |
| Chi-square of  -2log(likelihood) test, *p*-value | 138.740  <0.0001 | 128.451  <0.0001 | 61.943  <0.0001 |
| **Training Sample #**  Accuracy,  Sensitivity, Specificity,  Precision,  and F1 Score | Accuracy:0.6396  Sensitivity:0.6635  Specificity:0.6120  Precision:0.6635  F1 Score:0.6635 | Accuracy:0.6396  Sensitivity:0.6494  Specificity:0.6258  Precision:0.7109  F1 Score:0.6787 | Accuracy:0.5425  Sensitivity:0.7273  Specificity:0.3608  Precision:0.5279  F1 Score:0.6118 |
| **Validation Sample #**  Accuracy,  Sensitivity, Specificity,  Precision,  and F1 Score | Accuracy:0.5700  Sensitivity:0.6346  Specificity:0.5000  Precision:0.5789  F1 Score:0.6055 | Accuracy:0.5500  Sensitivity:0.6250  Specificity:0.4808  Precision:0.5263  F1 Score:0.5714 | Accuracy:0.5900  Sensitivity:0.6429  Specificity:0.5227  Precision:0.6316  F1 Score:0.6372 |
| Area under the ROC curve | 0.703 | 0.699 | 0.655 |

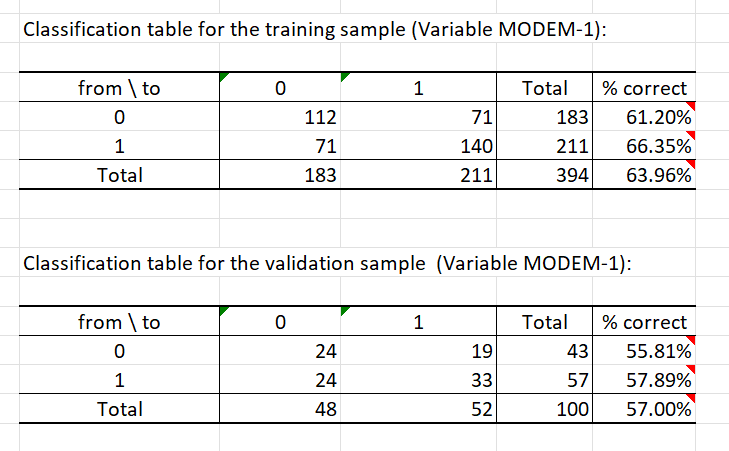
#Remark: Your group only needs to calculate the performance measures (accuracy, sensitivity, …) for your BEST model.

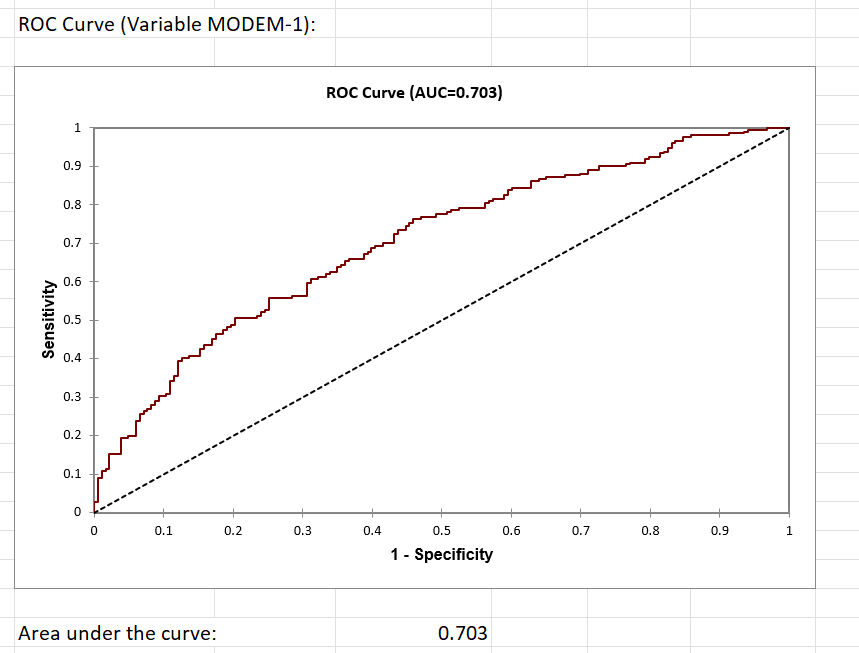
1. Result of our best model:

Of the three models listed in our report, our best model is model 1. Model 1 shows a significant P-value and high percentages on sensitivity, significance, and accuracy in both the training and validation samples. In model 1’s validation sample, precision was high at 0.5789 and in the training sample at 0.6635. Although not the highest out of all three models, model 1 shows a high area under the ROC curve at 0.703, the highest of all three models. This shows that through highlighting family employment status (FES) as a binary variable, we can surmise that as a family's employment status relates to more individuals bringing home money, their household income rises. Incorporating the numerical values of property value (VALP), the categorical variable of family employment status (FES) and the year the property was built (YBL) helps create a better calculation of the response variable household income (HINCP).

The F1 score in each model is relatively similar, however the F1 score in model 1’s training sample is 0.6635 and 0.6055 in the validation sample. Each F1 score furthers the accuracy of the model when compared to other models in large. Looking at model 1 in a broad sense, the accuracy of 0.6635,F1 score of 0.6635, and a significant P-value at <0.0001 lead us to believe that model 1 is our best model to prove family earning status is an accurate predictor of household income.







1. Data story: Please see step 7 in the instruction part. (1-2 paragraphs) （Marina）

The overall importance of model 1 is that it has a significant p-value and the validation sample showed a high precision. We looked at family employment status as a binary variable and thought about how they related to numerical values of property value, categorical variables family employment status and the year the property was built in order to calculate the response variable household income. In using binary logistic regression, we had two possible outcomes. However, there still was a large data set for getting higher accuracy in model one, even though it was significant enough it was not the most significant out of all our regression models.