**Bushra Hamdan Alghamdi**

**HAP 823**

**Assignment 1**

1. **Summarize the data.**

 We have a binomial outcome which is the success of the CIT or the Failure of it (its effectiveness as a treatment for antidepressant) satisfactory therapeutic response to CIT. The logistic regression analysis uses the concept of probabilities and log odds with cut-off probability 0.5.

We used logistic regression because we have one or more explanatory variables to predict a binary outcome.

**Multiple regression Summary output:**

|  |
| --- |
| SUMMARY OUTPUT |
|  |  |
| *Regression Statistics* |
| Multiple R | 0.191726969 |
| R Square | 0.03675923 |
| Adjusted R Square | 0.028715466 |
| Standard Error | 0.481846766 |
| Observations | 1933 |

1. **Select a set of variables and construct a logistic regression model to predict success of CIT. (The complete construction is in SPSS file)**

 There are a lot of automated methods that can choose what are the most effective variables in the final model. The three most commonly used methods are backward elimination, forward selection, and stepwise selection. However, it’s approved that the use of automated variables selection methods results in many problems one of them is that it results in estimated standard errors that are biased low.

For me to complete the assignment I’ve read the study protocol, and I also asked a family physician about which variables to include. Also as a first step to understand the variables relation to each other I did bivariate correlation to see the correlation between the success of CIT and each of the other variables. Based on the correlation table there was a strong positive correlation between CIT success and years of schooling completed and family history of depression. Also there were a strong negative correlation between CIT success years of first MDE, months in current episode, is recurrent, family history of bipolar, generalized anxiety diagnosis, panic diagnosis, social phobia diagnosis, obsessive compulsive diagnosis, alcohol abuse diagnosis, drug abuse diagnosis, hypochondriasis diagnosis, somatoform diagnosis, bulimia nervosa diagnosis, QIDS total score andHRSD-17. Only Age at first MDE was not significant.

So I chose a list of candidate variables that I will examine for their association with response to CIT. These variables include demographic characteristics (age and gender, race), years of schooling completed, age at first MDE, Years since first MDE, months in current episode, is recurrent, family hx depression, family hx bipolar, generalizes anxiety diagnosis, panic diagnosis, social phobia diagnosis, obsessive compulsive diagnosis, alcohol abuse diagnosis, drug abuse diagnosis, hypochon driasis diagnosis, somatoform m diagnosis, bulimia nervosa diagnosis, QIDS total score and HRDS-17.

1. **Check assumptions of the model through visual plots, including:**
* **the residuals of the model are nearly normal,**

We can see from the plot the assumption that the residuals are normally distributed is **violated**.



* **the variability of the residuals is nearly constant**

Because we have a binary model for an independent variable which takes 0,1 values, we have 2 lines here.



* **the residuals are independent**

From the plot we can see that residuals are **dependent**: negative correlation.



* **each variable is linearly related to the outcome**



I plotted each of the independent variable against the dependent variable to check this. (Complete charts for each variable in SPSS file)

Unlike ordinary linear regression, logistic regression is used for predicting binary dependent variables rather than a continuous outcome, so the assumptions of linearity is **violated**.

1. **Describe what predicts success of CIT.**

If we looked at CIT Actual success data and compared it to our model prediction values (predicted group (PGR\_1)) I plotted the frequency to see the different on the chart and the table. 1170 of the cases had a fail response to CIT, and 763 had a success response to CIT from total of 1933 cases. The same result obtained after I constructed the logistic regression in Block 0 (Classification table).

|  |
| --- |
| **CIT\_Success** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Fail | 1170 | 60.5 | 60.5 | 60.5 |
| Success | 763 | 39.5 | 39.5 | 100.0 |
| Total | 1933 | 100.0 | 100.0 |  |



**Prediction process:**

First I did the ROC which Shows the trade off in sensitivity and specificity for all possible thresholds.

A measure of goodness-of-fit often used to evaluate the fit of a logistic regression model is based on the simultaneous measure of sensitivity (True positive) and specificity (True negative) for all possible cutoff points. The area under the curve is .574 with 95% confidence interval. Also, the area under the curve is significantly different from 0.5 since p-value is .000 meaning that the logistic regression classifies the group significantly better than by chance. The graph shows that there is scope for improvement in terms of the predictive power of the model but the fitted model is still adequate (since a portion of the curve lies above the reference line).



|  |
| --- |
| **Area Under the Curve** |
| Test Result Variable(s): Predicted group  |
| Area | Std. Errora | Asymptotic Sig.b | Asymptotic 95% Confidence Interval |
| Lower Bound | Upper Bound |
| .601 | .013 | .000 | .574 | .627 |
| The test result variable(s): Predicted group has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased. |
| a. Under the nonparametric assumption |
| b. Null hypothesis: true area = 0.5 |

|  |
| --- |
| **Case Processing Summary** |
| Unweighted Casesa | N | Percent |
| Selected Cases | Included in Analysis | 1933 | 100.0 |
| Missing Cases | 0 | .0 |
| Total | 1933 | 100.0 |
| Unselected Cases | 0 | .0 |
| Total | 1933 | 100.0 |
| a. If weight is in effect, see classification table for the total number of cases. |

Block 0:

|  |
| --- |
| **Classification Tablea,b** |
|  | Observed | Predicted |
|  | CIT\_Success | Percentage Correct |
|  | Fail | Success |
| Step 0 | CIT\_Success | Fail | 1170 | 0 | 100.0 |
| Success | 763 | 0 | .0 |
| Overall Percentage |  |  | 60.5 |
| a. Constant is included in the model. |
| b. The cut value is .500 |

This is can be considered as our null hypothesis. We can see from the classification table that model predicts correct value in 60.5% of the cases. 1170 of the 1933 would fail (the CIT response), and 763 of the 1933 cases would have a successful response to the CIT, without any predicted variables involved in the model.

|  |
| --- |
| **Variables not in the Equation** |
|  | Score | df | Sig. |
| Step 0 | Variables | Gender(1) | .694 | 1 | .405 |
| Race | 14.042 | 6 | .029 |
| Race(1) | 2.775 | 1 | .096 |
| Race(2) | 8.238 | 1 | .004 |
| Race(3) | 1.495 | 1 | .221 |
| Race(4) | .457 | 1 | .499 |
| Race(5) | .013 | 1 | .909 |
| Race(6) | .652 | 1 | .419 |
| Years\_of\_schooling\_completed | 26.440 | 1 | .000 |
| Age\_at\_First\_MDE | 1.654 | 1 | .198 |
| Years\_Since\_First\_MDE | 6.083 | 1 | .014 |
| Months\_in\_current\_episode | 6.460 | 1 | .011 |
| Is\_Recurrent | .635 | 1 | .425 |
| Family\_Hx\_Depression | 5.519 | 1 | .019 |
| Family\_Hx\_Bipolar | .021 | 1 | .884 |
| Generalized\_Anxiety\_Diagnosis | 10.644 | 1 | .001 |
| Panic\_Diagnosis | 23.659 | 1 | .000 |
| Social\_Phobia\_Diagnosis | 17.019 | 1 | .000 |
| Obsessive\_Compulsive\_Diagnosis | 9.804 | 1 | .002 |
| Alcohol\_Abuse\_Diagnosis | 3.811 | 1 | .051 |
| Drug Abuse Diagnosis | 3.414 | 1 | .065 |
| Hypochondriasis\_Diagnosis | 7.439 | 1 | .006 |
| Somatoform\_Diagnosis | 9.464 | 1 | .002 |
| Bulimia\_Nervosa\_Diagnosis | 4.789 | 1 | .029 |
| QIDS\_totalscore\_CRC | 65.776 | 1 | .000 |
| HRSD\_17\_ROA | 54.875 | 1 | .000 |
| Overall Statistics | 136.113 | 25 | .000 |

The variables that are not in the equation which is the explanatory variables, it shows us if it’s not used in the model how strongly they will be able to create significant model. We can see that most of the variables have a p-value that is less than 0.05, that’s mean they are good predictors individually for the response to CIT.

1. **Describe how well the model predicts response to CIT.**

Block1:

|  |
| --- |
| **Omnibus Tests of Model Coefficients** |
|  | Chi-square | df | Sig. |
| Step 1 | Step | 144.346 | 25 | .000 |
| Block | 144.346 | 25 | .000 |
| Model | 144.346 | 25 | .000 |

Here in Block 1 from the output, we can see that the significant level is less than 0.05 and that means we have a **significant model** and the variables are **good predictors**.

|  |
| --- |
| **Hosmer and Lemeshow Test** |
| Step | Chi-square | df | Sig. |
| 1 | 8.070 | 8 | .427 |

The hosmer and lemeshow test gives us an idea about how good our model is. Here we want our p-value to be greater than 0.05, and yes it is which indicate we have a good model.

|  |
| --- |
| **Contingency Table for Hosmer and Lemeshow Test** |
|  | CIT\_Success = Fail | CIT\_Success = Success | Total |
| Observed | Expected | Observed | Expected |
| Step 1 | 1 | 161 | 161.258 | 32 | 31.742 | 193 |
| 2 | 144 | 144.514 | 49 | 48.486 | 193 |
| 3 | 129 | 134.690 | 64 | 58.310 | 193 |
| 4 | 126 | 127.250 | 67 | 65.750 | 193 |
| 5 | 127 | 119.758 | 66 | 73.242 | 193 |
| 6 | 117 | 112.821 | 76 | 80.179 | 193 |
| 7 | 113 | 105.571 | 80 | 87.429 | 193 |
| 8 | 105 | 98.120 | 88 | 94.880 | 193 |
| 9 | 79 | 89.930 | 114 | 103.070 | 193 |
| 10 | 69 | 76.089 | 127 | 119.911 | 196 |

Also in this table we will understand how good is our model in predicting the success of CIT. We will look at the last step (10). We can see that the observed number of success is 127 and our model were able to predict about 120 of those actual successful response to CIT (the closer those 2 numbers the better the model). So, our model is a good model because it was able to predict 120 out of 127.

|  |
| --- |
| **Classification Tablea** |
|  | Observed | Predicted |
|  | CIT\_Success | Percentage Correct |
|  | Fail | Success |
| Step 1 | CIT\_Success | Fail | 990 | 180 | 84.6 |
| Success | 492 | 271 | 35.5 |
| Overall Percentage |  |  | 65.2 |
| 1. The cut value is .500
 |

The classification shows us how good our model was in predicting the actual outcomes. Our model was able to predict 65.2% of the categories (65.2% of the outcomes was correctly predicted by our model.

|  |
| --- |
| **Variables in the Equation** |
|  | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | Gender(1) | .125 | .104 | 1.437 | 1 | .231 | 1.133 |
| Race |  |  | 5.138 | 6 | .526 |  |
| Race(1) | .440 | .463 | .904 | 1 | .342 | 1.553 |
| Race(2) | -.216 | .141 | 2.329 | 1 | .127 | .806 |
| Race(3) | -.362 | .273 | 1.761 | 1 | .184 | .696 |
| Race(4) | -.216 | .556 | .150 | 1 | .698 | .806 |
| Race(5) | .247 | .763 | .104 | 1 | .747 | 1.280 |
| Race(6) | -20.915 | 40192.970 | .000 | 1 | 1.000 | .000 |
| Years\_of\_schooling\_completed | .043 | .016 | 7.422 | 1 | .006 | 1.044 |
| Age\_at\_First\_MDE | -.001 | .004 | .095 | 1 | .758 | .999 |
| Years\_Since\_First\_MDE | -.005 | .004 | 1.125 | 1 | .289 | .995 |
| Months\_in\_current\_episode | -.002 | .001 | 2.999 | 1 | .083 | .998 |
| Is\_Recurrent | -.086 | .121 | .506 | 1 | .477 | .917 |
| Family\_Hx\_Depression | .223 | .102 | 4.743 | 1 | .029 | 1.250 |
| Family\_Hx\_Bipolar | -.030 | .172 | .031 | 1 | .861 | .970 |
| Generalized\_Anxiety\_Diagnosis | .056 | .142 | .155 | 1 | .694 | 1.058 |
| Panic\_Diagnosis | -.296 | .186 | 2.525 | 1 | .112 | .744 |
| Social\_Phobia\_Diagnosis | -.197 | .120 | 2.680 | 1 | .102 | .822 |
| Obsessive\_Compulsive\_Diagnosis | .056 | .171 | .107 | 1 | .743 | 1.058 |
| Alcohol\_Abuse\_Diagnosis | -.199 | .166 | 1.446 | 1 | .229 | .819 |
| Drug Abuse Diagnosis | -.151 | .218 | .478 | 1 | .489 | .860 |
| Hypochondriasis\_Diagnosis | -.321 | .308 | 1.090 | 1 | .296 | .725 |
| Somatoform\_Diagnosis | -.705 | .504 | 1.954 | 1 | .162 | .494 |
| Bulimia\_Nervosa\_Diagnosis | -.161 | .162 | .989 | 1 | .320 | .851 |
| QIDS\_totalscore\_CRC | -.077 | .016 | 23.703 | 1 | .000 | .926 |
| HRSD\_17\_ROA | -.027 | .007 | 15.649 | 1 | .000 | .973 |
| Constant | .918 | .412 | 4.963 | 1 | .026 | 2.505 |
| a. Variable(s) entered on step 1: Gender, Race, Years\_of\_schooling\_completed, Age\_at\_First\_MDE, Years\_Since\_First\_MDE, Months\_in\_current\_episode, Is\_Recurrent, Family\_Hx\_Depression, Family\_Hx\_Bipolar, Generalized\_Anxiety\_Diagnosis, Panic\_Diagnosis, Social\_Phobia\_Diagnosis, Obsessive\_Compulsive\_Diagnosis, Alcohol\_Abuse\_Diagnosis, Drug Abuse Diagnosis, Hypochondriasis\_Diagnosis, Somatoform\_Diagnosis, Bulimia\_Nervosa\_Diagnosis, QIDS\_totalscore\_CRC, HRSD\_17\_ROA. |

From the above table we can see that values like, years of schooling completed has a significant value and also family history of depression, QIDS total score and HRSD-17 which indicated that those variables the model (Exp(B)>1 indicates a positive correlation, and a negative of less). Variables like age at first MDE and family history of bipolar are not significant values which indicates that those variables do not improve the model.



The above graph shows that quite a lot of cases are actually in the middle area of the plot, i.e. the model is predicting a probability of around 0.5 (50:50 chance) that CIT *response* will be achieved.