**Binary logistic regression**

*I carried out a data analysis on cognitive failure of groups of quarry factory workers resulting from their exposure to noise generated at four quarry factories. The affected workers were quarry machine operators. The failure was carried out based on their measured defective hearing threshold at various frequencies. The failure was measured on a scale that assessed the degrees of the failure into five scale measurements. A set of workers, which consisted of clerical and administrative workers at the factories, was used to validate the model. I was charged with prescribing and employing a model that should assign the quarry machine operators into two categories of cognitive state – a category of those with cognitive failure and the other category without it – based on the scale measurements. I prescribed and employed binary logistic regression model. A snapshot of variable transformation, data coding, model classification and prediction classification of a set of workers for model validation is as follows:*

**Method for Cognitive Failure Evaluation**

Cognitive failure, as a dependent variable for evaluation of workers cognitive ability thought to have been caused by organisation of their work environment and inherent risk factors, is based on Likert’s five-point scale. The scale measures workers’ responses to questions testing their cognitive efficiency with rating levels: *very often, quite often, occasionally, very rarely and never*. Each option is scored on interval values ranging from 5 to 1 as shown in the table below:

|  |  |
| --- | --- |
| **Scale measurement** | **Score** |
| Very often | 5 |
| Quiet often | 4 |
| Occasionally | 3 |
| Very rarely | 2 |
| Never | 1 |

The scale consisted of twenty-five questions. Each respondent was scored on the twenty-five questions. The highest value a respondent could score was 125 (25 multiplied 5) and the lowest 25 (25 multiplied by 1). Each respondent’s score was adjusted to reflect the overall scale score based on the values of the scale response rating, namely 5, 4, 3, 2 and 1.

For the logistic regression used for cognitive failure modelling, those respondents with approximated scores of 5, 4 and 3 were assessed to possess sound cognitive ability and coded with 0s and those with scores of 2 and 1 (those with 40% or less) were assessed to suffer cognitive failure and coded 1s. The coding was done in R language algorithm.

**Results of Binary Logistic Regression of Cognitive Failure and Age of Worker, Years of Exposure and Noise Level**

**Table 1: Hosmer-Lemeshow Test of Model Fit**

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Chi-square** | **Df** | **Sig.** |
| 1 | 1.813 | 8 | 0.986 |

The Hosmer-Lemeshow statistics suggests a good fit, if the significance value is greater than 0.05. Since the significance value (0.986) is greater than 0.05, the logistic regression model fits the data adequately.

**Table 2: The Contingency Table for Hosmer-Lemeshow Test**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Step** | **Cognitive Failure = No** | | **Cognitive Failure = Yes** | | **Total** |
| **Observed** | **Expected** | **Observed** | **Expected** |
| 1 | 20 | 20.000 | 0 | .000 | 20 |
| 2 | 20 | 20.000 | 0 | .000 | 20 |
| 3 | 20 | 19.999 | 0 | .001 | 20 |
| 4 | 20 | 19.997 | 0 | .003 | 20 |
| 5 | 20 | 19.986 | 0 | .014 | 20 |
| 6 | 20 | 19.918 | 0 | .082 | 20 |
| 7 | 19 | 19.557 | 1 | .443 | 20 |
| 8 | 19 | 17.855 | 1 | 2.145 | 20 |
| 9 | 9 | 10.055 | 11 | 9.945 | 20 |
| 10 | 2 | 1.633 | 19 | 19.367 | 21 |

**Table 3: Variable Not Included in the Equation at Step Zero**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Step** | **Variable** | **Score** | **Df** | **Sig.** |
| 0 | Hearing loss  Age of worker  Years of exposure  Noise level | 98.681 | 1 | .000 |
| 73.239 | 1 | .000 |
| 74.215 | 1 | .000 |
| 9.952 | 1 | .002 |
| Overall Statistics | 101.279 | 4 | .000 |

The method begins with model that does not include any of the predictors (table 3) and it only includes the intercept.

**Table 4: R Statistics**

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **-2 Log likelihood** | **Cox & Snell R Square** | **Nagelkerke R Square** |
| 1 | 54.143 | .455 | .780 |

In linear regression model the coefficient of determination R square summarises proportion of dependent variable that is accounted for by the independent variable, with larger value indicating higher proportion of explanation of dependent variable by the model to a maximum of 1 (100%). For regression model with categorical dependent variable, such as logistic regression used, it is impossible to compute a single R squared value that includes all its properties in the linear regression model. The only feasible way is computation of Cox and Snell R and Nagelkerke’s R square approximations presented in table 4 above. Cox and Snell’s R squared (1989) is based on the log likelihood for model compared to the log likelihood for a baseline model. However, with category input relationship it has a theoretical maximum value of less than 1, even for perfect relationship. Nagelkerke’s R square value is adjusted value of Cox and Snell R squared to cover the full range from 0 to 1. Cox and Snell and Nagelkerke R squared values are 0.455 and 0.780 respectively. This model has as the independent variables (the factors) – age of worker, years of exposure and noise level. Using Nagelkerke’s R squared for full range of 0 to 1, hearing loss, age of worker, years of exposure and noise level explains 78% variation in cognitive failure; that is the three variables are accountable for 78% of the cognitive failure of workers, while 22% are explained by unidentified variables not included in the modelling.

For the purpose of hypothesis test, the null hypothesis is rejected and the alternative upheld. Upholding the alternative implies that hearing loss, age of worker, years of exposure and noise level can predict cognitive failure of workers under study.

**Table 5: Classification of Selected Cases of Respondents**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Observed** | | **Predicted** | | |  | |
| **Selected Cases** | | |
| **Cognitive failure** | | **Percentage Correct** | 36  0  3  2 | Percentage Correct |
| No | Yes |
| 1 | Cognitive failure | No | 164 | 5 | 97.0 | 94.4 |
| Yes | 6 | 26 | 81.3 | 80.0  92.7 |
| Overall Percentage |  |  |  | 94.5 |

Classification table 5 shows practical results of using logistic regression model. For each respondent (case) the predicted response is positive if the predicted value is more than the cut-off point, which is 0.5. Cells on the diagonal are correct predictions. Cells off the diagonal are incorrect predictions. Thus, of number of 201 cases used to create the model 26 of 32 who had cognitive failure are classified correctly; 164 of 169 of cases with no cognitive failure are classified correctly; and generally, 94.5% of the cases have been classified correctly.

**Table 6: Classification of Unselected Cases of Respondents**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Step | Observed | | **Predicted** | | |  | |
| **Unselected Cases** | | |
| **Category of cognitive failure** | | **Percentage Correct** | 36  0  3  2 | Percentage Correct |
| No | Yes |
| 1 | Cognitive failure | No | 36 | 1 | 97.3 | 94.4 |
| Yes | 1 | 3 | 75.0 | 80.0  92.7 |
| Overall Percentage |  |  |  | 95.1 |

Table 6 above shows the results of the classification of cases that were used to validate the model. These cases were not subjected to the noise level in the environment of the machine operation workers. Out of the 44 cases only 39 of them were selected by the software procedure due to missing values. Based on the model 36 out of 37 who had no cognitive failure are correctly classified as having none and 3 out of 4 with cognitive failure are correctly classified. On the whole, 95.1% of the cases have been classified correctly.