**Group G Assignment Report**

**How do demographic factors such as gender, address, schoolsup, etc., influence a student’s academic performance?**

**Introduction**

The data set that we are working with provides information based on the academic performance of students. Some of the attributes included in the dataset are gender, travel time, failures, etc. In the data set there is an attribute failure which gives us the history of the student’s academic performance. We concluded that this is a Case Control Study, as the case control study is also retrospective.

**AIM of doing the study based on Students Dataset**

Our intention with this study is to find factors that contributed to a student passing. We will do that by performing a test of association which will help us to determine if there’s a relationship between the attributes in the dataset and the variable of interest. Next, we will fit a statistical mode which will provide us with the intuitive visualization that will help us to identify the relationship between variables and make a prediction. We will find the best fitting model by fitting a reduced model, performing a goodness of fit test on our final model and checking for any influential observations. Lastly, we will interpret the results of our final model by commenting on the odds ratios and their confidence intervals and use the final model to determine the predicted probability of the event of interest occurring for the 3 observations given in our data set.

**Part one: Student Performance Data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Median** | **Mode** | **Variance** | **Standard Deviation** |
| Age | 16.74422 | 17.0000 | 17.0000 | 1.48386 | 1.21814 |
| Failures | 0.221880 | 0.00000 | 0.00000 | 0.35193 | 0.59324 |
| TravelTime | 1.568567 | 1.00000 | 1.00000 | 0.56049 | 0.74866 |
| SstudyTime | 1.930663 | 2.00000 | 2.00000 | 0.68809 | 0.82951 |

These measures describe the distribution, spread, and central tendency of the data, with standard deviation specifically indicating the average variability or spread of data from the mean.

Variables identified according to their respective measurement of scale.

|  |  |  |  |
| --- | --- | --- | --- |
| **Nominal** | **Ordinal** | **Interval** | **Ratio** |
| Gender | Go\_out | Age | Fam\_size |
| Adress | Free Time | Failures | Travel\_time |
| Schoolsup | Fam\_rel |  | Study time |
| Activities |  |  |  |
| Nursery |  |  |  |

**Interpreting the basic summary measures of the quantitative variables**

* **Age**
  + **Mean**: is 16.744 and it indicates the average age of all the students that are in the study.
  + **Median**: is 17 which means that 50% of the students younger than 17 years and 50% of them are older than 17 years.
  + **Mode**: 17 years is the number age that occurs most frequently within the age group of students.
  + **Standard Deviation**: is 1.218, it’s a measure of how spread out the data points are from the mean. Since it’s a small standard deviation than it indicates that the data points are closely clustered around the mean.
  + **Variance**: is 1.483, since it’s rather a small value it indicates that the ages of students are close to the mean.
  + **Range:** is 7, it’s the spread of the student’s ages from the highest age group to the lowest age group.
  + **Interquartile Range**: is 2, it’s a small value which indicates that the middle values cluster more tightly.
* **Failures:**
  + **Mean**: is 0.221880 and it indicates the average number of failures for students in the study.
  + **Median**: is 0.00, many values within the failure dataset are close to zero and very few values are one and two.
  + **Mode**: 0 occurs most frequently within the age group of students.
  + **Standard Deviation**: is 0.59324, it’s a measure of how spread out the data points are from the mean. Since it’s a small standard deviation than it indicates that the data points are closely clustered around the mean.
  + **Variance**: is 0.35193, since it’s rather a small value it indicates that the ages of students are close to the mean.
  + **Range:** is 3, it’s the spread of the student’s ages from the highest age group to the lowest age group.
  + **Interquartile Range**: is 0, it’s a small value which indicates that the middle values cluster more tightly.

**Part two: Test of Association**

**Relevant Categorical Variables of interest**

* Gender
* Address
* Famsize
* Schoolsup
* Activities
* Nursery

**Response variable of interest**

* Final\_grade

**Test of Association**

* **Gender and Final Grade**

**H0:** There is no association between gender and final grade.

**H1**: There is an association between gender and final grade.

Since the **p-value = 0.0151** of Chi-square test is < 0.05 we reject the null hypothesis therefore there an assocation between gender and final grade.

* **Address and Final Grade**

**H0:** There is no association between address and final grade.

**H1**: There is an association between address and final grade.

Since the **p-value = 0.0003** of Chi-square test is < 0.05 we reject the null hypothesis therefore there an assocation between address and final grade.

* **Schoolsup and Final Grade**

**H0:** There is no association between schoolsup and final grade.

**H1**: There is an association between schoolsup and final grade.

Since the **p-value = 0.5772** of Chi-square test is > 0.05 we do not reject the null hypothesis therefore there is no assocation between schoolsup and final grade.

* **Activities and Final Grade**

**H0:** There is no association between activities and final grade.

**H1**: There is an association between activities and final grade.

Since the **p-value = 0.1473** of Chi-square test is > 0.05 we do not reject the null hypothesis therefore there is no assocation between activities and final grade.

* **Nursery and Final Grade**

**H0**: There is no association between nursery and final grade.

**H1**: There is an association between nursery and final grade.

Since the **p-value = 0.6683** of Chi-square test is > 0.05 we do not reject the null hypothesis therefore there is no assocation between nursery and final grade.

* **Famsize and Final grade**

**H0:** There is no association between famsize and final grade.

**H1:** There is an association between the famsize and final grade.

Since the **p-value = 0.8107** of Chi-square test is > 0.05 we do not reject the null hypothesis therefore there is no assocation between famsize and final grade.

**Part three: Fitting a statistical Model**

The model that we fitted included all the relevant categorical explanatory variables as well as the quantitative explanatory variables in the data. And the results we used to compare the significance of the categorical explanatory variables in the data, were taken from the Wald Chi-square p-value. We used Wald test because it’s useful when you need to consider the association of multiple predictors with outcome together.

|  |  |  |
| --- | --- | --- |
| **Variable** | **2-way test of association**  **(Chi-square p-values)** | **Logistic Regression**  **(Wald Chi-square p-values)** |
| Gender | 0.0133 | 0.0833 |
| Adress | 0.0002 | 0.0059 |
| Schoolsup | 0.5108 | 0.1024 |
| Activities | 0.1411 | 0.2034 |
| Nursery | 0.6451 | 0.5764 |
| Famsize | 0.8107 | 0.6606 |

Based on the p-values from the Wald Chi-square p-value in the Logistic Regression:

**H0:** BJ = 0 vs **H1:** BJ ≠ 0

* Gender is not significant since the Wald’s Chi-square p-value > 0.05 therefore, we do not reject the null hypothesis hence Gender is not significant.
* Address is significant since the Wald’s Chi-square p-value < 0.05 therefore, we reject the null hypothesis hence address in significant.
* Schoolsup is significant since the Wald’s Chi-square p-value > 0.05 therefore, we do not reject the null hypothesis hence schoolsup is significant.
* Activities is not significant since the Wald’s Chi-square p-value > 0.05 therefore, we do not reject the null hypothesis hence activities is significant.
* Nursery is not significant since the Wald’s Chi-square p-value > 0.05 therefore, we do not reject the null hypothesis hence nursery is significant.
* Famsize is not significant since the Wald’s Chi-square p-value > 0.05 therefore, we do not reject the null hypothesis hence famsize is significant.

The difference in p-values comes from the fact that a two-way chi-square test measures association between two categorical variables in isolation, whereas Wald's chi-square test in logistic regression evaluates the significance of a variable within the context of a broader, multivariate model.

**Part four: Finding the best fitting model**

In addition to the model that we used on part three of this assessment. The first fitted reduced model was fitted by removing all the variables that were found to not have an association with the variable of interest. The second model included all the relevant categorical explanatory variables, which are gender, address, schoolsup, activities, nursery and famsize. From the three models we had, the best fitting model was selected based on comparing the Akaike’s Information Criteria (AIC) and c-statistic of the models. A lower AIC tells us that the model best fits the data better compared to a higher AIC. For the c-statistic it provides insight into the performance of a classification model, specifically in terms of its ability to discriminate between positive and negative outcomes. High c-value (closer to 1.0) indicates good discrimination, more concordant airs and model reliability. Low c-value (close to 0.5 or below) indicates poor discrimination, more dis-concordant pairs and need for improvement.

|  |  |  |
| --- | --- | --- |
| **MODELS** | **AIC** | **C-STATISTIC** |
| ORIGINAL MODEL | 669.212 | 0.784 |
| 1ST REDUCED MODEL | 788.532 | 0.583 |
| 2ND REDUCED MODEL | 785.736 | 0.621 |

From the results of the AIC and C-statistic we can see that the original model has lowest AIC and highest C-statistic value which is an indication that this model best fits the data and can be relied on to make accurate predictions, which is important in decision making context.

**Goodness of Fit test**

The goodness of fit test helps with determining whether the model adequately describes the data. The Hosmer-Lemeshow and Deviance statistics are used when testing the goodness of fit test. The Hosmer-Lemeshow statistic helps to determine if the logistic regression model adequately describes the observed data by comparing predicted probabilities with actual outcomes in groups. For the Hosmer-Lemeshow a low p-value indicates a significant difference between observed and expected values, suggesting poor model fit and a high p-value suggests that the model fits the data well. Deviance Statistic is a measure used to assess how well a logistic regression model fits the observed data, it is a measure of the discrepancy between the observed and predicted values. Lower deviance indicates a better fit. The hypothesis used for performing goodness of fit test are:

**H0**: The fitted model is adequate

**H1**: The fitted model is not adequate

Hosmer and Lemeshow Goodness-of-Fit Test:Since the p-value = 0.0671 > 0.05 we do not reject the null hypothesis. Therefore, the fitted model is adequate.

**𝐻0**: The fitted model does not improve significantly over the null model

**𝐻1**: The fitted model provides a significantly better fit than the null model.

Deviance and Pearson Goodness-of-Fit Statistics:Since value/df = 1.0800 and is closer to one, it indicates a better fit; we reject the hull hypothesis and conclude that the model provides a significantly better fit than the null model.

**Influential observations in the best fitted model**

**Part five: Interpreting the result**

**H0**: ¥ = 1 vs **H1**: ¥ ≠ 1

* **Gender (F vs. M)**: The odds ratio of 1.437 suggests that females have a higher likelihood of passing compared to males, but the confidence interval (0.953; 2.169) includes 1, indicating this result is not statistically significant.
* **Address (Urban vs. Rural)**: An odds ratio of 1.827 indicates that being in an urban area significantly increases the likelihood of passing, and the confidence interval (1.186; 2.805) does not include 1, making it statistically significant.
* **Schoolsup (Yes vs. No)**: The odds ratio of 0.593 suggests a lower likelihood of passing with school support, but the confidence interval (0.320 ;1.127) includes 1, indicating this result is not statistically significant.
* **Activities (Yes vs. No)**: The odds ratio of 1.291 shows a higher likelihood of passing with extracurricular activities, but the confidence interval (0.872;1.916) includes 1, meaning it’s not statistically significant.
* **Nursery (Yes vs. No)**: The odds ratio of 0.868 does not indicate a significant effect, as the confidence interval (0.522;1.414) includes 1.
* **Family Size (3 or less vs. More than 3)**: An odds ratio of 0.910 suggests a lower likelihood, but it is not statistically significant since the confidence interval includes 1: (0.597;1.394).
* **Age**: An odds ratio of 0.933 suggests a lower likelihood of passing, but the confidence interval (0.785;1.109) includes 1, indicating this result is not statistically significant.
* **Failure**: An odds ratio of 0.150 indicates that past failures significantly lower the likelihood of passing, and the confidence interval (0.088;0.242) does not include 1, making it statistically significant.
* **Famrel**: The odds ratio of 1.065 shows a higher likelihood of passing, but the confidence interval (0.870;1.299) includes 1, meaning it’s not statistically significant.
* **TravelTime**: The odds ratio of 0893 shows a lower likelihood of passing, but the confidence interval (0.689;1.165) includes 1, meaning it’s not statistically significant.
* **StudyTime**: The odds ratio of 1.504 shows a higher likelihood of passing, but the confidence interval (1.170;1.954) es not doinclude 1, meaning it’s statistically significant.

**Determining the predicted probability of the event of interest occurring for the first 3 observations given in the data set.**

The first 3 observation in the dataset are gender, address and schoolsup. In general, by looking at the predicted probability for both females and males whether they live in rural areas or not and whether they have schoolsup or not, the probability of passing their final grade decreases as the age increases.

**Conclusion**

By summarizing everything that has been done from performing a two-way test of association and to creating models and doing interpretations we can conclude that not all the variables in the data set have an effect of whether a student passes of fail their final grade.

**SAS Syntax**

**Part 1**

proc import datafile="/home/u63784591/STAT305/Assignment/dataset.xlsx"

out=StudentsData

dbms=xlsx

replace;

getnames = yes;

sheet="student-por";

run;

proc freq data=StudentsData;

tables gender address schoolsup activities nursery Final\_grade goout freetime famrel;

run;

proc univariate data = StudentsData;

var age failures traveltime studytime;

run;

**Part 2**

proc import datafile="/home/u63784591/STAT305/Assignment/dataset.xlsx"

out=Gender

dbms=xlsx

replace;

getnames = yes;

sheet="Gender";

run;

proc import datafile="/home/u63784591/STAT305/Assignment/dataset.xlsx"

out=Address

dbms=xlsx

replace;

getnames = yes;

sheet="Address";

run;

proc import datafile="/home/u63784591/STAT305/Assignment/dataset.xlsx"

out=Schoolsup

dbms=xlsx

replace;

getnames = yes;

sheet="Schoolsup";

run**;**

proc import datafile="/home/u63784591/STAT305/Assignment/dataset.xlsx"

out=Nursery

dbms=xlsx

replace;

getnames = yes;

sheet="Nursery";

run;

proc import datafile="/home/u63784591/STAT305/Assignment/dataset.xlsx"

out=Activities

dbms=xlsx

replace;

getnames = yes;

sheet="Activities";

run;

title'Gender as Categorical variable of interest and Final\_grade as the response variable';

proc freq data=Gender order=data;

tables gender\*Final\_grade/ nocol nopct chisq cmh;

run;

title'Address as Categorical variable of interest and Final\_grade as the response variable';

proc freq data= Address order=data;

tables address\*Final\_grade/ nocol nopct chisq cmh;

run;

title'Schoolsup as Categorical variable of interest and Final\_grade as the response variable';

proc freq data= Schoolsup order=data;

tables schoolsup\*Final\_grade/ nocol nopct chisq cmh;

run;

title'Activities as Categorical variable of interest and Final\_grade as the response variable';

proc freq data= Activities order=data;

tables activities\*Final\_grade/ nocol nopct chisq cmh;

run;

title'Nursery as Categorical variable of interest and Final\_grade as the response variable';

proc freq data= Nursery order=data;

tables nursery\*Final\_grade/ nocol nopct chisq cmh;

run;

**Part 3 and 4**

proc import datafile="/home/u63784591/STAT305/Assignment/dataset.xlsx"

out=StudentsData

dbms=xlsx

replace;

getnames = yes;

sheet="student-por";

run;

/\* Fitted using all the variables from the dataset \*/

/\* Best fitted model \*/

ods graphics on;

proc logistic data=StudentsData plots=all order=data;

class final\_grade gender(ref="M") address(ref="Rural") schoolsup(ref="no") activities(ref="no") nursery(ref="no") famsize(ref="More than 3")/

param=reference;

model final\_grade (event="Pass") = gender address schoolsup activities nursery famsize age failures famrel traveltime studytime /

aggregate scale=none clparm=wald clodds=pl rsquare lackfit;

\*\*output out=results predicted=pihat dfbetas=\_all\_ difchisq=chisq reschi=pearsonr resdev=g2res;

run;

ods graphics off;

/\*Fitted using significant variables\*/

ods graphics on;

proc logistic data=StudentsData plots=all order=data;

class final\_grade address(ref="Rural") schoolsup(ref="no")/ param=reference;

model final\_grade (event="Pass") = address schoolsup /aggregate scale=none clparm=wald clodds=pl rsquare lackfit influence;

run;

ods graphics off;

/\*Fitted using relevant categorical explanatory variables\*/

ods graphics on;

proc logistic data=StudentsData plots=all order=data;

class final\_grade gender(ref="M") address(ref="Rural") schoolsup(ref="no") activities(ref="no") nursery(ref="no") famsize(ref="More than 3");

model final\_grade (event="Pass") = gender address schoolsup activities nursery famsize /aggregate scale=none clparm=wald clodds=pl rsquare lackfit influence;

run;

ods graphics off;