# Multiple Regression model

Nitin Gupta - x19112033[1] and Christopher Herrera Magana - x19127723[2]

[1]National College of Ireland, Cloud Computing: Research Methods
[1]Total number of words ≈2693

## I. OBJECTIVE

The objective of this analysis is to perform multiple regression analysis on the dataset, average housing prices in Ireland and generate a regression model to check if the prices depends on the independent variables like year, house age, distance from transport hub, number of stores, latitude and latitude. The model can be used to predict the future price variation accordingly. Multiple regression is similar to linear regression which is used to predict the value of dependent variable based on one or more independent variable. Multiple regression helps to understand the overall fit of the model based on variance and respective contribution of each independent variable in the total prediction model.

## II. STATEMENT OF THE PROBLEM: QUESTION

Does transaction year, house age, distance from nearest Luas, bus or Dart station, number of convenience store, latitude, longitude affect the price of a house in Ireland?

## III. MULTIPLE REGRESSION MODEL LOOKS LIKE:

$$Y = \alpha + \beta 1X1 + \beta 2X2 + ... + \beta kXk \qquad (1)$$

Y = dependent variable
X1,X2,X3 = independent variables
$\alpha = constant$
$\beta 1, \beta 2, \beta 3 = coefficients$

## IV. VARIABLES

The dataset has been taken from t https://data.gov.ie/ site. This data is available from 2012 to 2018. Data cleaning has been done before proceeding to the analysis. Like null values has been replaced with the mean of the independent variable column. The dataset consists of 6 independent variables (transaction year, house age, distance from nearest Luas, bus, Dart station, number of convenience store, latitude, longitude) and 1 dependent variable (price).

## V. HYPOTHESIS

The Significance value of 0.05 has been used in all the results.

1. Null Hypothesis – There is no relationship between the independent and dependent variables means the coefficients values are 0.

$$H_0 : \beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6 = 0$$

2. Alternative Hypothesis - There is significant relationship between independent and dependent variables.

$$H_1 : \beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6 \neq 0$$

We will proceed with the multiple regression analysis and see if there is enough evident to reject the Null hypothesis. But we need to check the assumptions required for the dataset to fulfil.
Y = Price
X = transaction year, house age, distance from nearest Transport (Luas, bus, Dart station), number of stores, latitude, longitude.

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Included | | Excluded | | Total | |
| | N | Percent | N | Percent | N | Percent |
| Price * Year | 414 | 100.0% | 0 | 0.0% | 414 | 100.0% |
| Price * house_age | 414 | 100.0% | 0 | 0.0% | 414 | 100.0% |
| Price * D_MRT | 414 | 100.0% | 0 | 0.0% | 414 | 100.0% |
| Price * CV | 414 | 100.0% | 0 | 0.0% | 414 | 100.0% |
| Price * Lat | 414 | 100.0% | 0 | 0.0% | 414 | 100.0% |
| Price * Lon | 414 | 100.0% | 0 | 0.0% | 414 | 100.0% |

Fig. 1. Case processing summary

1. Test of Normality SPSS test has been used to test the normality of the dataset. Figure 2 shows the dependent variable (price) is normally distributed. The same result can be seen in the histogram plot. The bell curve shows the data to be normal in nature.

**Tests of Normality**

| | | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | Year | Statistic | df | Sig. | Statistic | df | Sig. |
| Price | 2012.00 | .065 | 95 | .200* | .984 | 95 | .283 |
| | 2013.00 | .056 | 149 | .200* | .989 | 149 | .294 |
| | 2014.00 | .113 | 16 | .200* | .956 | 16 | .598 |
| | 2015.00 | .134 | 13 | .200* | .937 | 13 | .417 |
| | 2016.00 | .096 | 23 | .200* | .964 | 23 | .546 |
| | 2017.00 | .170 | 17 | .200* | .932 | 17 | .231 |
| | 2018.00 | .130 | 26 | .200* | .971 | 26 | .653 |
| | 2019.00 | .099 | 75 | .065 | .866 | 75 | .000 |

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Fig. 2. Test of normality

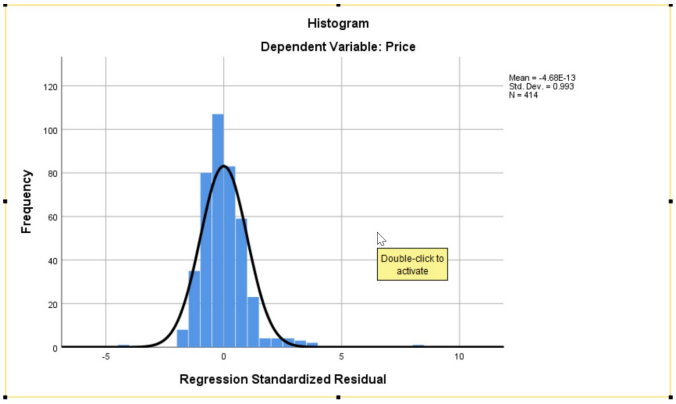


Fig. 3. Histogram dataset

2. Test of Homogeneity Another assumption check required for multiple regression is tests of homogeneity of data. The test performed using SPSS tool and the results shown in figure 4 shows the significance value pvalue = 0.886 (¿ 0.05). This satisfies the requirement to proceed with our analysis.

**Test of Homogeneity of Variance**

| | | Levene Statistic | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| Price | Based on Mean | .021 | 1 | 412 | .886 |
| | Based on Median | .036 | 1 | 412 | .849 |
| | Based on Median and with adjusted df | .036 | 1 | 406.844 | .849 |
| | Based on trimmed mean | .013 | 1 | 412 | .909 |

Fig. 4. Test of normality

## VI. ANALYSIS

Regression Equation:

$$Price = \alpha + \beta1(Year) + \beta2(houseage) + \beta3(distancefromnearesttransport) + \beta4(numberofstores) + \beta5(latitude)\beta6(latitude)$$

The coefficients value defines the relationship between dependent and independent variables. If the coefficients value is 0 this means, there the independent variables do not have any effect on the dependent variable [1]. If the coefficient values are positive, which means the slope is positive and going upward. This depicts that with the increase in independent variables, corresponding dependent variable (price) will also increase. It's difficult to know the exact value of these coefficients in real time scenarios.

To perform multiple regression analysis, IBM SPSS tool has been used. Figure 5 shows the model summary in which

**Model Summary[b]**

| | | | | | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | R Square Change | F Change | df1 | df2 | Sig. F Change | Durbin-Watson |
| 1 | .763[a] | .582 | .576 | 8.85767 | .582 | 94.591 | 6 | 407 | .000 | 2.153 |

a. Predictors: (Constant), Lon, Year, house_age, Lat, CV, D_MRT

b. Dependent Variable: Price

Fig. 5. Model summary

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 44528.847 | 6 | 7421.474 | 94.591 | .000[b] |
| | Residual | 31932.531 | 407 | 78.458 | | |
| | Total | 76461.378 | 413 | | | |

a. Dependent Variable: Price

b. Predictors: (Constant), Lon, Year, house_age, Lat, CV, D_MRT

Fig. 6. Analysis of Variance

The Analysis of variance or ANOVA section shows how much variation is there in the prices given in the dataset. The total Sum of squares value (76461.378) shows how much our dependent variable (price) is spread out. Out of 76461.378, 44528.847 explains by our regression model using the 6 variables (degree of freedom) used in our model and 31932.531 is the residual. Residual mean square (78.458) shows on an average how much each observation in our dataset missing our prediction model. It's quite low which shows our model is goodfit. R square value = 44528.847/76461.378 = 0.58237 - This is correct as can be seen in figure above in model summary. This means that 58.2% of the variation

in price is dependent on our X variables. There's still 41.8% of residual is unexplained means price variation cannot be explained using our X variables. The F (6, 407) = 94.591 which is quite significant. The p value (at 5% of significance level) is 0.000 hence this shows that there is very less probability that the improvements shown in the model is due to random chance and not dependent on the X variables. This leads to the rejection of our Null hypothesis and concludes that there is likely significant relationship between one or all independent variables and dependent variable. The ANOVA results only shows if any of the independent variable is significant in the model but to find out exactly which variables, we need to look at figure 4.

Figure 4 shows the constant value -14437.101 (value of alpha in the regression equation) and respective coefficient value of the variables.

$$Price = 5303.530 + 0.089(Year) - 0.268(houseage) - 0.004(distancefromnearesttransport) + 1.164(numberofstores) + 238.702(latitude) - 6.532(longitude)$$

Interpretation of results:

• The above equation explains that for every additional year, the price will increase by 0.089 for the house considering the other independent variables unchanged.

• Similarly, for every additional year of house age, the price will decrease by 0.268 considering the other variables constant

• For every additional meter distance from nearest transport, the price will decrease by 0.004 considering other variables constant.

• For every additional number of stores, the price will increase by 1.164 of the houses considering other variables constant.

At first, upon looking at these interpretations, latitude appears to be highly significant in predicting the price of the house and longitude has negative relation with the price. But we look at the corresponding p values of all the variables, the p value for all the variables is less than 0.05 except longitude. Hence, we can assume that longitude has no significance in depicting the price of the house hence it can be neglected from the regression equation.

The final equation comes out to be :

$$Price = -5303.530 + 0.089(Year) - 0.268(houseage) - 0.004(distancefromnearesttransport) + 1.164(numberofstores) + 238.702(latitude)$$

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. | Correlations Zero-order | Correlations Partial | Correlations Part | Collinearity Statistics Tolerance | Collinearity Statistics VIF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | (Constant) | -5303.530 | 6252.326 | | -.848 | .397 | | | | | |
| | Year | .089 | .166 | .017 | .536 | .059 | .015 | .027 | .017 | .993 | 1.008 |
| | house_age | -.268 | .039 | -.224 | -6.846 | .000 | -.211 | -.321 | -.222 | .982 | 1.018 |
| | D_MRT | -.004 | .001 | -.393 | -5.853 | .000 | -.674 | -.279 | -.190 | .233 | 4.292 |
| | CV | 1.164 | .190 | .252 | 6.114 | .000 | .571 | .290 | .198 | .620 | 1.614 |
| | Lat | 238.702 | 45.021 | .218 | 5.302 | .000 | .546 | .254 | .172 | .624 | 1.601 |
| | Lon | -6.532 | 49.249 | -.007 | -.133 | .895 | .523 | -.007 | -.004 | .341 | 2.931 |

a. Dependent Variable: Price

Fig. 7. Coefficients

The corresponding t values shown in figure tells that convenience stores and latitude has higher positive significance and house age and distance from transportation has negative significance in depicting the dependent variable. In the start of our analysis, we assume the null hypothesis that the coefficients are zero but upon looking the above figure the coefficients turn out to be far away from zero. Now this can happen by chance in the sample taken out of whole population. The p values tell us how likely that is. The p value gives us the probability of these coefficients occurring by chance.

## VII. Conclusion

For year: If our Null Hypothesis is true ($\beta = 0$), the chance of us getting our sample data as extreme as this ($\beta 1 = 0.89$), is 5.9%. For House_age: If our Null Hypothesis is true ($\beta = 0$), the chance of us getting our sample data as extreme as this ($\beta 2 = -0.268$), is almost 0%. For Transportation distance: If our Null Hypothesis is true ($\beta = 0$), the chance of us getting our sample data as extreme as this ($\beta 3 = -0.004$), is almost 0%. For Number of stores: If our Null Hypothesis is true ($\beta = 0$), the chance of us getting our sample data as extreme as this ($\beta 4 = 1.164$), is almost 0% For Latitude: If our Null Hypothesis is true ($\beta = 0$), the chance of us getting our sample data as extreme as this ($\beta 5 = 138.702$), is almost 0% For Longitude: If our Null Hypothesis is true ($\beta = 0$), the chance of us getting our sample data as extreme as this ($\beta 6 = -6.532$), is 89.5%. it means that there is highly likely a chance of getting this coefficient value nonzero. Hence, we reject this value.

## VIII. Logistic Regression

### A. Introduction

Amid this Covid-19 disease, there is uncertainty regarding the Online teaching courses. Some authors claim that the present teaching model does not provide the same results, and it affects students' performance [2] . Moreover, international students complain that there is no reduction in virtual classes, causing many students to cancel their plans to take an abroad course. Other students opinion is about the interval of hours connected to online classes and the distractions that this might bring due to other students or the Internet connection [**?**], [3].

### B. Problem Statement

This paper's main objective is to predict and convey a logistic regression analysis of the association among international English students based on studying hours effectively, scores, and final grades obtained during this Covid-19 disease. This statistical model is expected to identify if the independent variable influences the dependable (binary) variable.

### C. Significance of the study

This investigation can provide a better insight into how students feel about online classes and what aspects could enhance to improve student's adaptation. Using a predictive logistic regression, we could create a model to determine the number of hours suitable for self-learning, reducing additional stress and burnout. Another example of this logistic regression analysis can be observed in the Irish Leaving Certificate(reference).

### D. Literature Review

No one ever imagined that a pandemic might occur and how it will affect the whole world. It is believed that the face-to-face teaching model was adequate for this era. However, Covid-19 disease changed the way we teach or learn. Some authors claimed that there is no significant difference between traditional and online classes [4]; However, there are some factors to consider in which students prefer one teaching module. Some students might struggle with online courses due to the lack of experience with that delivery model. Other might like the self-learning and keep their learning speed.

The present Online teaching model does not have the correct infrastructure to maintain such unbalance classes [5], [6]. Despite the efforts to adapt the courses Online, one of the main factors affecting the teaching delivery model is due to the lack of technology skills [7]. Moreover, the educative model has not changed. The use of presentations or videos does not look attractive to students who invest money just for that experience. Nevertheless, this obstacle might help to improve the teaching and learning experience [8].

There has been some research in the use of Virtual Reality (VR) and Augmented Reality (AR) to enhance online classes [9], [10]. This type of idea could revolutionize the way we learn, and we teach. But there are some challenges to adopt this technology. The cost of VR headsets could be a burden for some students and schools. In addition, the transformation of the old program to a new infrastructure. To the best of our knowledge, many authors wonder what would be the direction for this post-Covid-19 online teaching. However, there is not research into predictive analysis on the student's affinity and performance.

### E. Source of the Data and Sample Selection

The data sets for this investigation were obtained from Kaggle [11]. Due to the limitations and some variables, we had to use a random number generator to adapt some of this dataset's categorical variables. The dataset for this statistic analysis contains two independent quantitative variables: score and time, in the case of the score, it is measure from 0 up to 10, and the time is measured in hours per week. Finally, the dependent binary variable is the Grade obtained in this case is 0 if the student fails or 1 if the student passes. Also, we selected 77 different students so we could check the error of prediction with small samples and big samples.

## IX. Methods, Analysis & Results

We the use of R programming language to analyze our independent and dependent variables, we can proceed with the statistical analysis. R code will be presented in the appendix section A.

### A. Descriptive Statistics

For this logistic regression there is not need to check the distribution of our sample, however there are some assumptions that need to be satisfy in order to proceed with the logistic regression model. The Fig 13 shows the summary of our dataset.

### B. Significance Level

As we are analyzing a binary logistic regression a significance level ($\alpha$) usually of 0.05 will indicate the risk of concluding that there exist a relationship between the variables when there is not association [12]. Therefore, in this logistic regression analysis we select 5% as our significance level(reference)

### C. Assumption ♯1:

Dependent variables should be measured on a dichotomous scale. In other words, it will have only two values True or False or Yes or No. In our dataset, we have the dependent variable (Grade), which it fulfills this assumption. In the appendix A section, there is a visual table representation of our dataset.

### D. Assumption ♯2:

Independent variables should be continuous either interval or ratio variable( ordinal or nominal). In this logistic statistic analysis, our dataset contains two continuous independent variables: score and hours. Hence, we satisfy this assumption.

### E. Assumption ♯3:

Independence of observation means that each student is only counted as one observation. Furthermore, the dependent variable should have mutually exclusive, which translates that if one thing happens, other things cannot occur. There is not enough information regarding the source of the dataset and how did they obtain the values. Also, we cleaned the dataset and made modifications to be able to proceed with this analysis. Hence we can consider this with the results of the study and model obtained.

### F. Assumption ♯4:

Linearity assumption states that the relationship between continuous predictor variables and the logit of the outcome should be approximately linear [13].

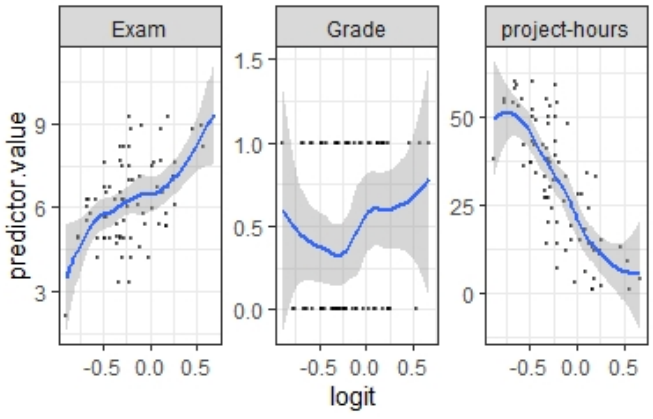


Fig. 8. Linearity Assumption Analysis

As we observe from the images above the variables Exam and project-hours shows quite a linearity association with the grade logit. Therefore we satisfy this condition.

### G. Assumption ♯5:

Influential values are data points that can influence the quality of the logistic regression model. Using Cook's distance values representation we can check which is the most significant value that could affect our logistic analysis.

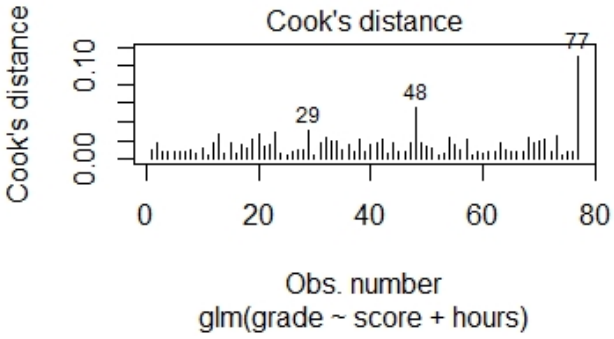


Fig. 9. Cook's distance analysis model

As we visualize the Fig (9) there are two values that could affect the interpretation of the logistic model.

Another important analysis is the standardized residuals in which we can compare the behaviour of the two independent variables with the dependent variable. Based on the results on the Fig (10) we can conclude that there is not influential observations in our data.

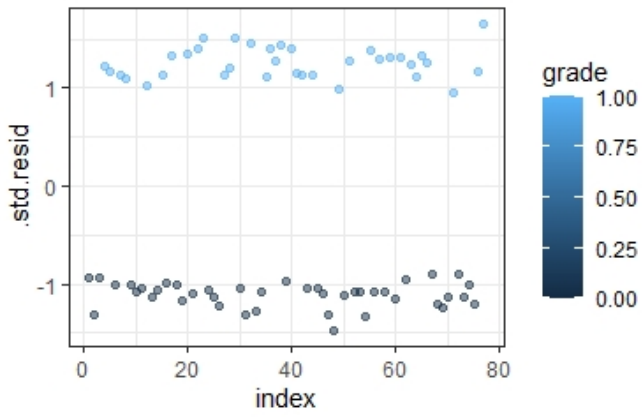


Fig. 10. Residuals model correlation

It is require to check that our sample follows and appropriate correlation as we observed from the Fig (11) the values are properly distributed.

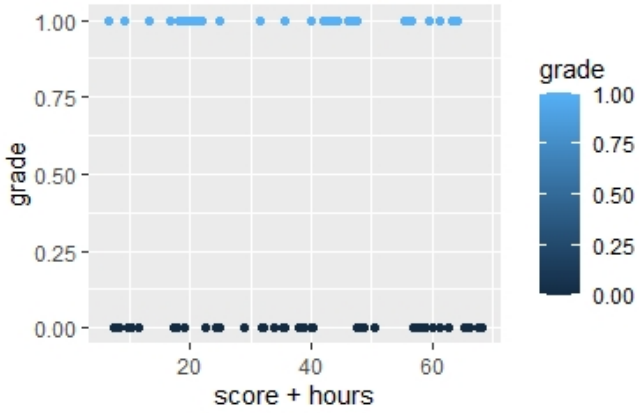


Fig. 11. Correlation

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4158  -1.0664  -0.9171   1.2049   1.5722

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.56364    1.01951  -0.553    0.580
score        0.14120    0.15681   0.900    0.368
hours       -0.01646    0.01341  -1.227    0.220

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 106.11  on 76  degrees of freedom
Residual deviance: 104.08  on 74  degrees of freedom
AIC: 110.08

Number of Fisher Scoring iterations: 4
```

Fig. 12. Logistic model results

### H. Logistic regression analysis

After validating the assumptions and confirming that our dataset fulfills them to a certain extent, we can now proceed to analyze the logistic regression model, interpretation, and equation. It is essential to clarify that the use of this statistics model is to predict a value based on prior observations or historical data [14].

*1) Logistic Regression model:*

$$\begin{aligned}\ln(\tfrac{P_g}{P_1}) &= \ln(\tfrac{P_g}{P_1}) + \beta_{g1}X_1 + \beta_{g2}X_2 + ...\beta_{gp}X_p \\ &= \ln(\tfrac{P_g}{P_1}) + XB_g\end{aligned} \tag{2}$$

Where $P_g$ is the probability that an individual value $X_1, X_2, ..., X_p$ is an outcome g. In other words,

$$P_g = \Pr(Y = g|X)$$

Usually $X_1 \equiv 1$ an intercept is included but might not be necessary. The quantities $P_1, P_2, ..., P_G$ represent the prior probabilities of outcome membership.(REFERENCE PDF) The regression coefficients $\beta_{11}, \beta_{12}, ..., \beta_{1p}$ for the reference value are set to zero. Usually, it is the most frequent value or a control outcome to which the other outcomes are to be compared [15].

Another important equation is the Log odds (logit) transformation the probability of success, $\pi$ :

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{3}$$

*2) Null Hypothesis:*

$$H_0 : \beta_i = 0$$

*3) Alternative Hypothesis:*

$$H_1 : \beta_i \neq 0$$

As we can observed the coefficient values are higher than our alpha value. The association is not statistically significant, we failed to reject the null hypothesis [12]. The cause of this results might depend of the model constructed and also the values of our dataset which are affecting the logistic regression model. At first glance the Fig (13) shows that the prediction is not accurate which is the reason that we are obtaining those results. The variables that are affecting this is the hours and the score which are not adequate with grade variable. As we previous mention the model requires a bigger dataset which can provide more accurate results.

| | grade | score | hours | .fitted | .se.fit | .resid | .hat | .sigma | .cooksd | .std.resid | index |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <int> |
| 1 | 0 | 6.3 | 60 | -0.662 | 0.457 | -0.912 | 0.0468 | 1.19 | 0.00886 | -0.934 | 1 |
| 2 | 0 | 7.1 | 12 | 0.241 | 0.380 | -1.28 | 0.0357 | 1.18 | 0.0163 | -1.30 | 2 |
| 3 | 0 | 6.28 | 59 | -0.649 | 0.446 | -0.917 | 0.0448 | 1.19 | 0.00855 | -0.938 | 3 |
| 4 | 1 | 5.75 | 19 | -0.0644 | 0.283 | 1.20 | 0.0201 | 1.19 | 0.00743 | 1.22 | 4 |
| 5 | 1 | 5.95 | 15 | 0.0296 | 0.311 | 1.16 | 0.0241 | 1.19 | 0.00819 | 1.18 | 5 |
| 6 | 0 | 6.7 | 52 | -0.473 | 0.364 | -0.984 | 0.0314 | 1.19 | 0.00694 | -1.00 | 6 |
| 7 | 1 | 6.55 | 14 | 0.131 | 0.330 | 1.12 | 0.0271 | 1.19 | 0.00837 | 1.14 | 7 |
| 8 | 1 | 7 | 11 | 0.244 | 0.383 | 1.08 | 0.0362 | 1.19 | 0.0102 | 1.10 | 8 |
| 9 | 0 | 4.88 | 34 | -0.435 | 0.330 | -0.999 | 0.0261 | 1.19 | 0.00593 | -1.01 | 9 |
| 10 | 0 | 4.2 | 20 | -0.300 | 0.405 | -1.05 | 0.0401 | 1.19 | 0.0107 | -1.07 | 10 |

Fig. 13. Logistic Model

The regression model model shows that there is low probability to pass the course, that is another finding that our dataset does not provide enough information to build a predictive model [16]. Also as the Receiver Operating Characteristic Curve illustrates that our data fits in the along th line means that our model does not provide good results.

$$logit(p) = -0.56364 + 0.14120(score) - 0.01646(hours) \tag{4}$$


Fig. 14. Logistic prediction model

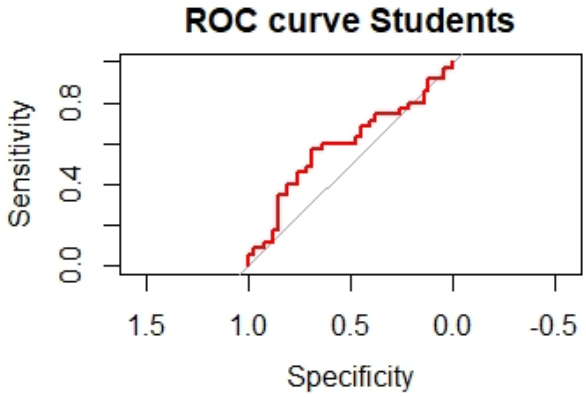


Fig. 15. Receiver Operating Characteristic Curve

## X. Conclusions

One of the challenges that we came across is the manipulation of the variables in the dataset and if the data source provide a good recollection which could cause bad interpretation on the logistic regression model. However, the logistic statistics analysis involved other aspects of probability which were useful to create a dummy model which is not accurate enough it gives some approximation how does student can pass the module.

Finally there are some gaps and complex interpretation to provide a better model, the future work for this is that we need to provide a better quality of data and also requires huge amount in order to predict the grades of the students using hours and previous scores. This statistics analysis provide us the tools to create a basic model.

## References

[1] "Anova test: Definition, types, examples - statistics how to." https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova. (Accessed on 07/08/2020).

[2] M. University, "Covid-19 mental health survey by maynooth university and trinity college finds high rates of anxiety — maynooth university." Available:https://www.maynoothuniversity.ie/news-events/covid-19-mental-health-survey-maynooth-university-and-trinity-college-finds-high-rates-anxiety.

[3] UNICEF, "'what will a return to school during the covid-19 pandemic look like?' — unicef." https://www.unicef.org/coronavirus/what-will-return-school-during-covid-19-pandemic-look. (Accessed on 07/08/2020).

[4] N. Kemp and R. Grieve, "Face-to-face or face-to-screen? undergraduates' opinions and test performance in classroom vs. online learning," *Frontiers in Psychology*, vol. 5, p. 1278, 2014. doi: 10.3389/fpsyg.2014.01278 JCR Impact Factor: (2.129).

[5] X. Zhu and J. Liu, "Education in and after covid-19: Immediate responses and long-term visions," *Postdigital Science and Education*, Apr 2020.

[6] "Teacher's virtual reality platform aims to improve student engagement." https://www.irishtimes.com/business/innovation/teacher-s-virtual-reality-platform-aims-to-improve-student-engagement-1.3754661. (Accessed on 07/08/2020).

[7] "How will the covid-19 pandemic impact the future of education? - wise." https://www.wise-qatar.org/how-will-the-covid-19-pandemic-impact-the-future-of-education/. (Accessed on 07/08/2020).

[8] "'what will a return to school during the covid-19 pandemic look like?' — unicef." https://www.unicef.org/coronavirus/what-will-return-school-during-covid-19-pandemic-look. (Accessed on 07/08/2020).

[9] "Vr in the classroom: The future of immersive education with ar/vr." https://edtechmagazine.com/k12/article/2019/08/arvr-k-12-schools-use-immersive-technology-assistive-learning-perfcon. (Accessed on 07/08/2020).

[10] "Frontiers — schooling beyond covid-19: An unevenly distributed future — education." https://www.frontiersin.org/articles/10.3389/feduc.2020.00082/full. (Accessed on 07/08/2020).

[11] B. Visser, "Aws spot pricing market — kaggle." https://www.kaggle.com/noqcks/aws-spot-pricing-market?select=eu-west-1.csv, 2017. (Accessed on 07/08/2020).

[12] r tutor, "Significance test for logistic regression — r tutorial." http://www.r-tutor.com/elementary-statistics/logistic-regression/significance-test-logistic-regression. (Accessed on 07/08/2020).

[13] "Logistic regression assumptions and diagnostics in r - articles - sthda." http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/. (Accessed on 07/08/2020).

[14] "Interpret the key results for binary logistic regression - minitab express." https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/interpret-the-results/key-results/#:~:text=A%20significance%20level%20of%200.05,there%20is%20no%20actual%20association.&text=If%20the%20p%2Dvalue%20is,response%20variable%20and%20the%20term. (Accessed on 07/08/2020).

[15] "Logistic regression." https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Logistic_Regression.pdf. (Accessed on 07/08/2020).

[16] "Logistic regression with r." http://r-statistics.co/Logistic-Regression-With-R.html. (Accessed on 07/08/2020).

[17] "Multiple logistic regression example." http://ismayc.github.io/teaching/sample_problems/multiple_logistic.html. (Accessed on 07/08/2020).

[18] J.-A. Baird, "Microsoft word - predictability report final3 _24jun14_." https://www.examinations.ie/about-us/Predictability-Overall-Report.pdf. (Accessed on 07/08/2020).
[19] C. O'Brien, "Leaving cert cancelled: Students who want written exam will have to wait." https://www.irishtimes.com/news/education/leaving-cert-cancelled-students-who-want-written-exam-will-have-to-wait-1.4247877. (Accessed on 07/08/2020).
[20] S. Bowers, "Leaving cert student told he is not eligible for predicted grade." https://www.irishtimes.com/news/education/leaving-cert-student-told-he-is-not-eligible-for-predicted-grade-1.4296785, 2020 July. (Accessed on 07/08/2020).
[21] "The rise of online learning during the covid-19 pandemic — world economic forum." https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/. (Accessed on 07/08/2020).
[22] minitab, "Interpret the key results for binary logistic regression - minitab express." https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/interpret-the-results/key-results/. (Accessed on 07/08/2020).
[23] S. Swaminathan, "Logistic regression — detailed overview - towards data science." https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc. (Accessed on 07/08/2020).
[24] "Multiple logistic regression - handbook of biological statistics." http://www.biostathandbook.com/multiplelogistic.html#:~:text=The%20main%20null%20hypothesis%20of,you%20would%20expect%20by%20chance. (Accessed on 07/08/2020).
[25] "How to perform a binomial logistic regression in spss statistics — laerd statistics." https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php. (Accessed on 07/08/2020).
[26] "Chapter 6 logistic regression — broadening your statistical horizons." https://bookdown.org/roback/bookdown-bysh/ch-logreg.html. (Accessed on 07/08/2020).
[27] J. Kim, "Teaching and learning after covid-19." https://www.insidehighered.com/digital-learning/blogs/learning-innovation/teaching-and-learning-after-covid-19. (Accessed on 07/08/2020).
[28] EURYDICE, "Focus on: What has the covid-19 crisis taught us about online teaching? — eurydice." https://eacea.ec.europa.eu/national-policies/eurydice/content/focus-what-has-covid-19-crisis-taught-us-about-online-teaching_en. (Accessed on 07/08/2020).
[29] "Five tips for moving teaching online as covid-19 takes hold." https://www.nature.com/articles/d41586-020-00896-7. (Accessed on 07/08/2020).
[30] W. Bao, "Covid-19 and online teaching in higher education: A case study of peking university," *Human Behavior and Emerging Technologies*, vol. 2, no. 2, pp. 113–115, 2020.
[31] "One-way anova in spss statistics - step-by-step procedure including testing of assumptions.." https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php. (Accessed on 07/08/2020).

## APPENDIX

### Dataset table

```
library(readxl)
info_spot <- read_excel("C:/Users/Chris/Desktop/
    Research Methods/regression/info spot.xlsx")
View(info_spot)
mydata <- info_spot
mydata
```

```
# A tibble: 77 x 4
   Exam `project-hours` Grade  prob
   <db1>           <db1> <db1> <db1>
 1  6.3                60     0 0.340
 2  7.1                12     0 0.560
 3  6.28               59     0 0.343
 4  5.75               19     1 0.484
 5  5.95               15     1 0.507
 6  6.7                52     0 0.384
 7  6.55               14     1 0.533
 8  7                  11     1 0.561
 9  4.88               34     0 0.393
10  4.2                20     0 0.426
# ... with 67 more rows
```

Fig. 16. Dataset Table

```
> describe(mydata)
              vars  n  mean    sd median trimmed   mad min   max range  skew kurtosis   se
Exam             1 77  6.28  1.53   6.27    6.27  1.74 2.1  9.25  7.15  0.01    -0.32 0.17
project-hours    2 77 30.96 17.82  33.00   31.14 25.20 1.0 60.00 59.00 -0.04    -1.25 2.03
Grade            3 77  0.45  0.50   0.00    0.44  0.00 0.0  1.00  1.00  0.18    -1.99 0.06
```

Fig. 17. Descriptive Statistics

```
str(mydata)
library(ggplot2)
ggplot(mydata, aes("Exam","Grade"))+
geom_point()
View(info_spot)
library(readxl)
info_spot <- read_excel("C:/Users/Chris/Desktop/
    Research Methods/regression/info spot.xlsx")
View(info_spot)
mydata <-info_spot
mydata
clear
ggplot(mydata, aes('Exam', 'Grade')) +
geom_point()
score <- mydata$Exam
grade <- mydata$Grade
hours <- mydata$`project-hours`
plot(score, jitter(grade, 0.15),pch= 19 +
xlab ="points", ylab= "result")
plot(score, jitter(grade, 0.15),pch= 19, +
xlab ="points", ylab= "result")
plot(score, jitter(grade,0.15), pch = 19, xlab="
    score", ylab="result")
plot(hours, jitter(grade,0.15), pch = 19, xlab="
    score", ylab="result")
plot(score + hours , jitter(grade,0.15), pch = 19,
    xlab="score", ylab="result")
model <- glm(grade~score+hours,binomial)
summary(model)
xv <- seq(min(score+hours), max(score+hours),0.01)
yv <- predict(model,list(score=xv),type="response")
yv <- predict(model,list(score+hours=xv),type="
    response")
yv <- predict(model,list(xv),type="response")
lines(xv,yv,col="red")
ggplot(myData, aes(score, grade)) +
geom_point()
ggplot(mydata, aes(score, grade)) +
geom_point()
ggplot(mydata, aes(score+hours, grade)) +
```

```
geom_point()
ggplot(mydata, aes(score+hours, grade)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE) +
coord_cartesian(ylim = c(0, 1))
ggplot(mydata, aes(score+hours, grade)) +
geom_point() +
geom_smooth(method = "glm", se = FALSE, method.args
    = list(family = "binomial"))
ggplot(mydata, aes(score, grade)) +
geom_point() +
geom_smooth(method = "glm", se = FALSE, method.args
    = list(family = "binomial"))
ggplot(mydata, aes(score, grade)) +
geom_point() +
geom_smooth(method = "glm", se = FALSE, method.args
    = list(family = "binomial"))
ggplot(mydata, aes(hours, grade)) +
geom_point() +
geom_smooth(method = "glm", se = FALSE, method.args
    = list(family = "binomial"))
ggplot(mydata, aes(hours, grade)) +
geom_point() +
geom_smooth(method = "glm", se = FALSE, method.args
    = list("binomial"))
ggplot(mydata, aes(hours, grade)) +
geom_point() +
geom_smooth(method = "glm", method.args = list("
    binomial"))
ggplot(mydata, aes(score, grade)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "glm", method.args = list(
    family = "binomial"))
mylogit <- glm(score ~ score + hours, data = mydata,
     family = "binomial")
mylogit <- glm(grade ~ score + hours, data = mydata,
     family = "binomial")
summary(mylogit)
confint(mylogit)
confint.default(mylogit)
exp(coef(mylogit))
plot_correlation(na.omit(mydata), maxcat = 5L)
coord_map(na.omit(mydata), maxcat = 5L)
plot_correlation(na.omit(mydata), maxcat = 5L)
ggplot(mydata, aes(score+hours, grade)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "glm", method.args = list(
    family = "binomial")) +
labs( title = "Logistic Regression Model",
x = "Score and Hours of Study",
y = "Probability of Passing the course")
qplot(x = score + hours, y = grade, color = grade,
    data = mydata, geom = "point")
summary(mylogit)
mylogit2 <- glm(grade ~ score, data = mydata, family
     = "binomial")
summary(mylogit2)
model
probabilities <- predict(model, type="response")
probabilities
predicted.classes <- ifelse(probabilities > 0.5, "
    pos", "neg")
head(predicted.classes)
predicted.classes
probabilities
mydata
mydata2 <- mydata2 %>%
mutate(logit = log(probabilities/(1-probabilities)))
     %>%
gather(key = "predictors", value = "predictor.value"
    , -logit)
mydata2 <- mydata
mydata2
mydata2 <- mydata2 %>%
mutate(logit = log(probabilities/(1-probabilities)))
     %>%
gather(key = "predictors", value = "predictor.value"
    , -logit)
mydata2
mydata2 <- mydata2 %>%
mutate(logit = log(probabilities/(1-probabilities)))
     %>%
gather(key = "predictors", value = "predictor.value"
    , -logit)
mydata2
mydata2 <- mydata2 +
mutate(logit = log(probabilities/(1-probabilities)))
     +
gather(key = "predictors", value = "predictor.value"
    , -logit)
mydata2<- mydata2%>%
mutate(logit = log(probabilities/(1-probabilities)))
     %>%
gather(key = "predictors", value = "predictor.value"
    , -logit)
install.packages("dplyr")
mydata2<- mydata2%>%
mutate(logit = log(probabilities/(1-probabilities)))
     %>%
gather(key = "predictors", value = "predictor.value"
    , -logit)
mutate(mtcars, displ_l = disp / 61.0237)
library(dplyr)
library(dbplyr)
mydata2<- mydata2%>%
mutate(logit = log(probabilities/(1-probabilities)))
     %>%
gather(key = "predictors", value = "predictor.value"
    , -logit)
install.packages("tidyr")
library(tidyverse)
mydata2<- mydata2%>%
mutate(logit = log(probabilities/(1-probabilities)))
     %>%
gather(key = "predictors", value = "predictor.value"
    , -logit)
ggplot(mydata2, aes(logit, predictor.value))+
geom_point(size = 0.5, alpha = 0.5) +
geom_smooth(method = "loess") +
theme_bw() +
facet_wrap(~predictors, scales = "free_y")
plot(model, which = 4, id.n = 3)
model.data <- augment(model) %>%
mutate(index = 1:n())
install.packages("msm")
library(broom)
model.data <- augment(model) %>%
mutate(index = 1:n())
model.data %>% top_n(3, .cooksd)
ggplot(model.data, aes(index, .std.resid)) +
geom_point(aes(color = diabetes), alpha = .5) +
theme_bw()
ggplot(model.data, aes(index, .std.resid)) +
geom_point(aes(color = grade), alpha = .5) +
theme_bw()
ggplot(model.data, aes(hours, grade)) +
geom_point(aes(color = grade), alpha = .5) +
theme_bw()
ggplot(model.data, aes(index, .std.resid)) +
geom_point(aes(color = grade), alpha = .5) +
theme_bw()
model.data %>%
filter(abs(.std.resid) > 3)
library(car)
car::vif(model)
xtabs(~grade + score, data = mydata)
xtabs(score+ hours ~ grade, data = mydata)
```

```
anova(model, test="Chisq")
install.packages("ROCR")
library(ROCR)
ggplot(mydata, aes(x=Rating, y=Recommended)) + geom_
    point() +
stat_smooth(method="glm", family="binomial", se=
    FALSE)
ggplot(mydata, aes(x=score+hours, y=grade)) + geom_
    point() +
stat_smooth(method="glm", family="binomial", se=
    FALSE)
ggplot(mydata, aes(x=score, y=grade)) + geom_point()
     +
stat_smooth(method="glm", family="binomial", se=
    FALSE)
desc(mydata)
probabilities
predicted.classes
roc.plot(observed = predicted.classes, Model.fit =
    Logis.Int.update)
install.packages("pROC")
roc.plot(observed = predicted.classes, Model.fit =
    Logis.Int.update)
library(pROC)
roc.plot(observed = predicted.classes, Model.fit =
    Logis.Int.update)
install.packages("plotROC")
library(plotROC)
roc.plot(observed = predicted.classes, Model.fit =
    Logis.Int.update)
detach("package:plotROC", unload = TRUE)
library(plotROC)
install.packages("ggROC")
library(ggROC)
plot.roc(observed = predicted.classes, Model.fit =
    Logis.Int.update)
ggroc(data = predicted.classes, bin = 0.02, roccol =
     "green", sp=19)
ggroc(data = mydata, bin = 0.02, roccol = "green",
    sp=19)
ggroc(data = mydata, bin = 0.02, roccol = "green",
    sp=19,output =roc.pdf )
ggroc(data = mydata, bin = 0.02, roccol = "green",
    sp=19,output =roc.jpeg )
ggroc(data = mydata, bin = 0.02, roccol = "green",
    sp=19,output ="roc.pdf")
probabilities
View(model.data)
ggroc(data =model , bin = 0.02, roccol = "green", sp
    =19,output ="roc1.pdf")
model
model.data
ggroc(data =model.data , bin = 0.02, roccol = "green
    ", sp=19,output ="roc1.pdf")
ggroc(data =model , bin = 0.02, roccol = "green", sp
    =19,output ="roc1.jpeg")
ggroc(data =mydata , bin = 0.02, roccol = "green",
    sp=19,output ="roc1.jpeg")
desc(mydata)
summary(mydata)
library(Hmisc)
install.packages("Hmisc")
library(Hmisc)
describe(mydata)
install.packages("psych")
library(psych)
describe(mydata)
model.data
mylogit
probabilities
mydata2
mydata
mydata2
```

```
mydata3 <mydata
mydata3 <- mydata
mydata3$prob = probabilities
mydata3
g <- roc(grade ~ prob, data = mydata3)
prob
prob <- probabilities
prob
g <- roc(grade ~ prob, data = mydata3)
g
plot(g)
ggplot(g)
plot(g, col=rainbow(7), main="ROC curve Students",
    xlab="Specificity",
ylab="Sensitivity")
abline(0, 1)
plot(g, col=rainbow(7), main="ROC curve Students",
    xlab="Specificity",
ylab="Sensitivity")
abline(0, 1)
plot(g, col=rainbow(7), main="ROC curve Students",
    xlab="Specificity",
ylab="Sensitivity")
model
summary(model)
mydata3
savehistory("C:/Users/Chris/Desktop/code in R.
    Rhistory")
```