**Linear Regression**

This is INDIVIDUAL assignment.

Guidelines:

1. Write your answer in R Markdown and answer below each question
2. Attach R Script along with your answer
3. Show RELEVANT R output in your answers (R Markdown will do this)
4. Follow the hint given after the question in [ ] brackets
5. Read the relevant topics as mentioned in the hint from Ken Black. You can refer any good article/blog from net also (but referring Ken Black is a must)
6. You are required to follow the word limits in stricter terms. Exceeding word limits will reduce your grades in this assignment.
7. You can take more space for your answers if needed.
8. This assignment would require a thorough/solid reading from book and net roughly of 20 hrs by each participant (not collective hrs of study!)

----------------------------------------------------------------------------------------------------------------

**Task**

Refer file **grades**.csv

The school principal wants to build a predictive model for predicting final for his consumption. As a principal he is very keen to have good scores by his students. He has given this data file to you with a request to suggest an appropriate model.

You are required to build at **least 4 models** with different sets of predictors (independent variables). Selection of sets of predictor/s is upon you. Different sets of predictors can be a single variable or more than one variable. However, selection of predictor/s should be based on some logic. For example, for predicting final score of students, roll number cannot be a logical predictor.

You will analyze all 4 models based on following points and recommend **the best** model to the Principal.

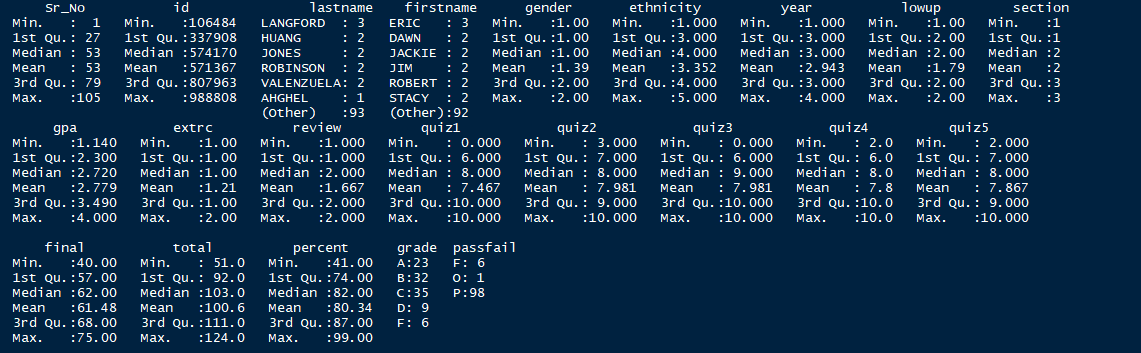
1. Describe data of response variable and predictors in terms of key summary statistics like mean, mode, median, standard deviation, range, skewness and kurtosis. Show histogram and box plots also for each variables. [hint: describe command in R]

Each variable to be explained in 30 words maximum.

**Analysis Response:**

A summary view of the dataset - grades.csv looks like as follows -





As can be noted for the output presented above –

1. The dataset includes a total of 22 variables of which the 14 variables (Sr\_No, id, lastname, firstname, gender, ethnicity, year, lowup, section, extrc, review, percent, grade, passfail) are categorical data.
2. The remaining 8 variables (*GPA, Quiz1, Quiz2, Quiz3, Quiz4, Quiz5, Final, Total*) appears to be non-categorical information/Contuinual readings representing a sample of records that can be used to build inferences about the population/derive some predictive measures and subsequently an attempt for recommendation of improvement plans can be established.
3. The variables GPA and Total seems to reflect a combined effect of rest other five quiz variables and doesn’t essentially enable a meaningful elaboration of real influencer for the overall goal. Thus, GPA and Total are the least contributors from statistical significance perspective and can be dropped from the analysis process.

**Know your Data:** A closure a look at each of the continuous variables to develop foundation for initial analysis is as follows -

|  |  |
| --- | --- |
| 1 | * describe(grades$gpa) * hist(grades$gpa, main = "Histogram of GPA", xlab = "GPA", ylab = "Frequency", col = "Blue") * boxplot(grades$gpa, main = "Box plot of GPA", xlab = "GPA", col = "Blue", horizontal = T) * stem(grades$gpa) |
|  | |

|  |  |
| --- | --- |
|  |  |
| 2 | * describe(grades$quiz1) * hist(grades$quiz1, main = "Histogram of quiz1", xlab = "Quiz1 Marks", ylab = "Marks", col = "Blue") * boxplot(grades$quiz1, main = "Box plot of Quiz1", xlab = "Quiz1", col = "Blue", horizontal = T) * stem(grades$quiz1) |
|  | |

|  |  |
| --- | --- |
|  |  |
| 3 | * describe(grades$quiz2) * hist(grades$quiz2, main = "Histogram of quiz2", xlab = "Quiz2 Marks", ylab = "Frequency", col = "Blue") * boxplot(grades$quiz2, main = "Box plot of Quiz2", xlab = "Quiz2", col = "Blue", horizontal = T) * stem(grades$quiz2) |
|  | |

|  |  |
| --- | --- |
|  |  |
| 4 | * describe(grades$quiz3) * hist(grades$quiz3, main = "Histogram of quiz3", xlab = "Quiz3 Marks", ylab = "Frequency", col = "Blue") * boxplot(grades$quiz3, main = "Box plot of Quiz3", xlab = "Quiz3", col = "Blue", horizontal = T) * stem(grades$quiz3) |
|  | |

|  |  |
| --- | --- |
|  |  |
| 5 | * describe(grades$quiz4) * hist(grades$quiz4, main = "Histogram of quiz4", xlab = "Quiz4 Marks", ylab = "Frequency", col = "Blue") * boxplot(grades$quiz4, main = "Box plot of Quiz4", xlab = "Quiz4", col = "Blue", horizontal = T) * stem(grades$quiz4) |
|  | |

|  |  |
| --- | --- |
|  |  |
| 6 | * describe(grades$quiz5) * hist(grades$quiz5, main = "Histogram of quiz5", xlab = "Quiz5 Marks", ylab = "Frequency", col = "Blue") * boxplot(grades$quiz5, main = "Box plot of Quiz5", xlab = "Quiz5", col = "Blue", horizontal = T) * stem(grades$quiz5) |
|  | |

|  |  |
| --- | --- |
|  |  |
| 7 | * describe(grades$final) * hist(grades$final, main = "Histogram of final", xlab = "Final Marks", ylab = "Frequencys", col = "Blue") * boxplot(grades$final, main = "Box plot of Final", xlab = "Fianl", col = "Blue", horizontal = T) * stem(grades$final) |
|  | |

1. How predictor/s is related to response variable (final)? [hint: first plat scatter diagram followed by correlation test]

Present diagram/s and correlations in the following space. Before diagrams explain relationship in 3 or 4 lines.

**Analysis Response:**

|  |
| --- |
| The co-relation test results of response variable *final* with anticipated list of predictors (quiz1 to quiz5) indicate a positive co-relation with moderate level of strength on an average –quiz3 being in strongest form (0.5611773) and thus qualifies as best predictor candidate among the set. The other 4 prospect predictors though in moderate form, have positive co-relation leading to a decision for inclusion in the scope of study.  Thus the finalized set of predictors are – *quiz1, quiz2, quiz3, quiz4, quiz5* for the response variable *final*  The following diagrams display a scatter plot and output of correlation test within the context of investigation – |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

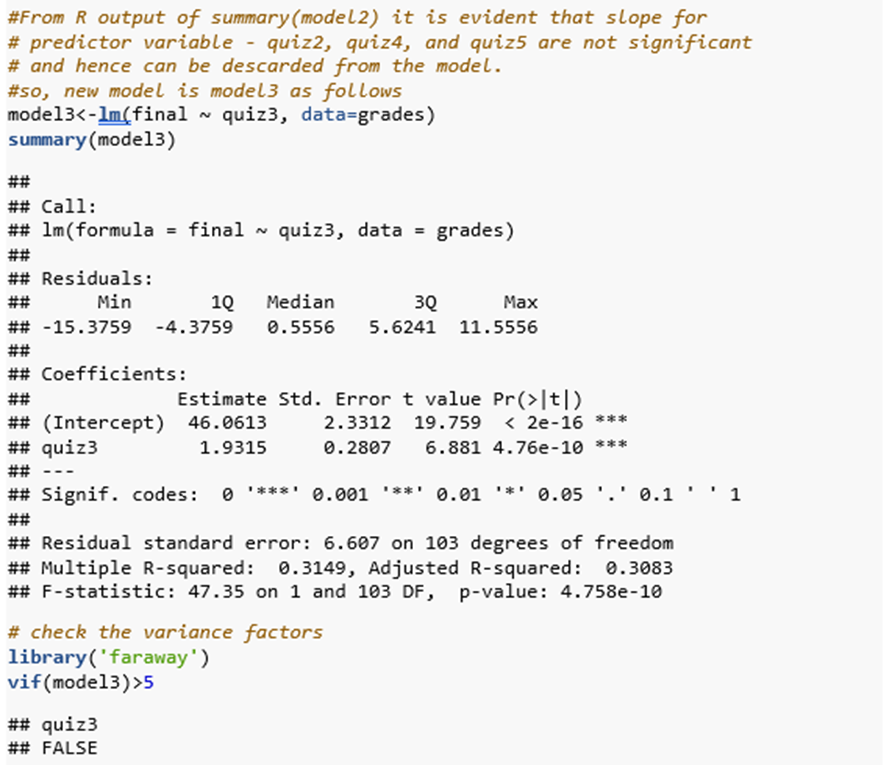
1. What are R Square and Adjusted R Square of your final model? Show R Output and explain in 3 or 4 lines. [hint: R Square and Adjusted R Square]

Explain the difference between R Square and Adjusted R Square. Which one is superior and why? Explain in maximum 4 lines.

**Analysis Response:**

The various models built as part of analysis exercise show that model-3 is the one delivering highest level of influence and hence is the best model with R Square value as 0.3149 and Adjusted R Square value as 0.3083.

|  |  |  |  |
| --- | --- | --- | --- |
| **##** | **Regression Models** | **R Square Value** | **Adj. R Square Value** |
| 1 | quiz1+quiz2+quiz3+quiz4+quiz5 | 0.3823 | 0.3512 |
| 2 | quiz2+quiz3+quiz4+quiz5 | 0.3808 | 0.356 |
| 3 | quiz3 | 0.3149 | 0.3083 |



The regression equation for model3 is -

|  |
| --- |
| Final at quiz3 = 46.061+1.931 quiz3 |

**R-Squared vs Adjusted R Squared:** R-squared measures the proportion of the variation in your dependent variable (final) explained by your independent variables (quiz3) for a linear regression model. However, the problem with R-squared is that it will either stay the same or increase with addition of more variables, even if they do not have any relationship with the output variables for example model2 in our case includes variable quiz2, quiz4 and quiz5 which are inflating the R Squared value without significant contribution to the acceleration in response variable (final).

Adjusted R-squared adjusts the statistic based on the number of independent variables in the model or in other words Adjusted R-square penalizes for adding variables which do not influence the response variable significantly - That is the desired property of a goodness-of-fit statistic and thus is superior to R Squared.

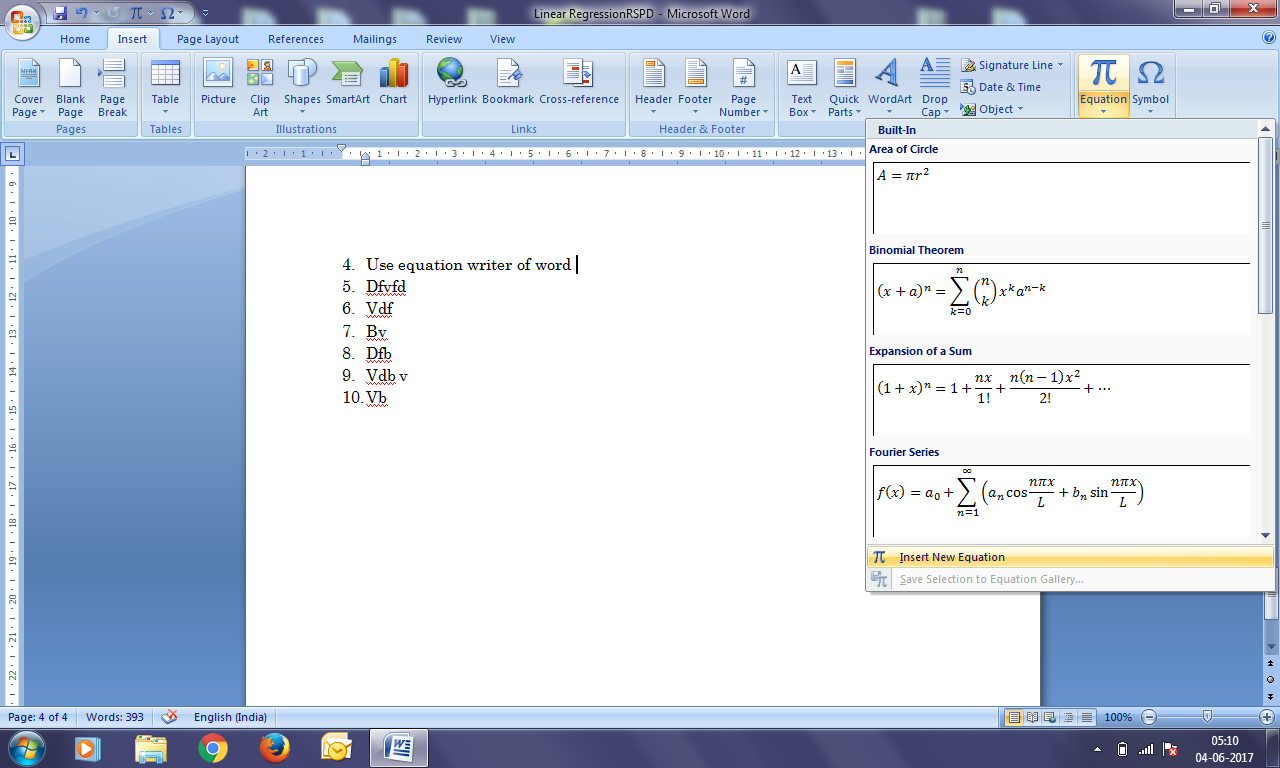
Typically, the more non-significant variables present into the model, higher the gap in R-squared and Adjusted R-squared is reflected.

1. How do you interpret significance value of *F*-statistics? Mention in 4 lines and show R Output. [Fitness of model]

**Analysis Response:**

|  |  |
| --- | --- |
| An F statistic is a value obtained from anova test or regression analysis to find out if the means between two populations are significantly different. It is used to identify if a group of variables are jointly significant - probability that the null hypothesis for the full model is true  *(i.e., that all regression coefficients are zero)*. |  |

1. Use equation writer of word [Insert 🡪 Equation 🡪 Insert New Equation and write Regression equation of the best model. Show R Output. [hint: refer summary of the model from R Output]



**Analysis Response:**

Write equation here 🡪

Final at quiz3 = 46.061+1.931 quiz3

Where –

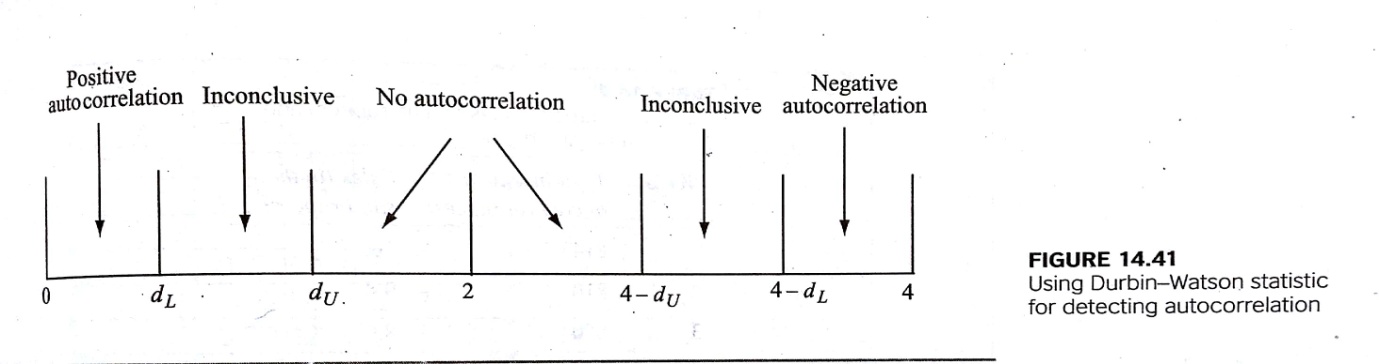
46.061 is the intercept

1.931 is the regression co-efficient

Quiz3 is the interval where a prediction is desired

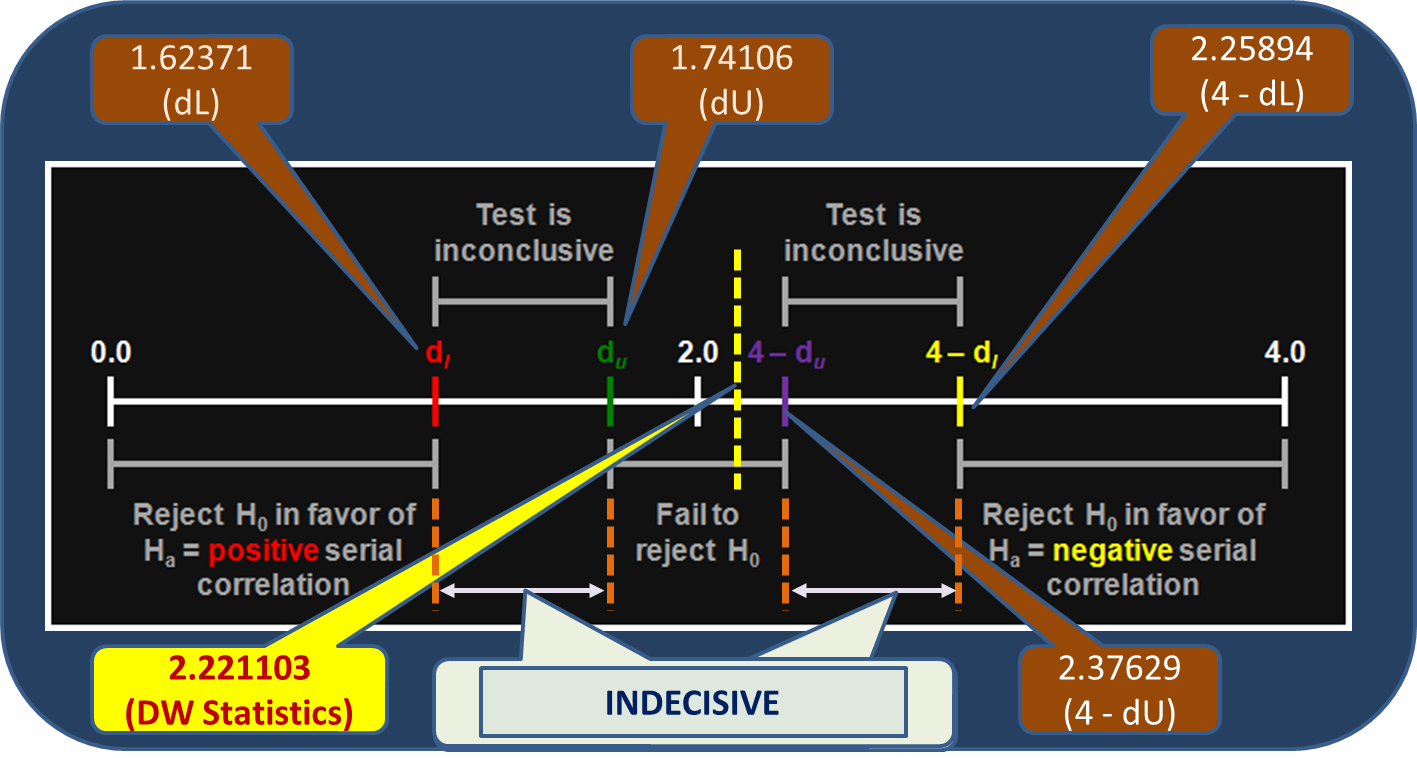
1. What is Durbin Watson Statistics of your model? How DWS is interpreted? Show how do you find dL and dU and design four boundaries in the sample diagram (SHARED in whatsapp group also). Maximum 5 lines. [hint: explore about Durbin Watson Statistics and table from internet. Table is used for finding dL and dU based on which you will design limits. You need to impose your DWS value in the diagram and decide about presence of autocorrelation]

Show R Output also.



**Analysis Response:**

|  |  |
| --- | --- |
| **Autocorrelation** (also known as serial co-relation) is a characteristic of data in which the correlation between the values of the same variables is based on related objects. It violates the assumption of instance independence, which underlies most of the conventional models.  In **statistics**, the **Durbin**–**Watson statistic** is a test **statistic** used to detect the presence of autocorrelation (a relationship between values separated from each other by a given time lag) in the residuals (prediction errors) from a regression analysis. It is named after James **Durbin** and Geoffrey **Watson**. Its value always remains between 0 to 4.  *A value of 2 means that there is no autocorrelation, 0 to <2 is positive autocorrelation and >2 to 4 is negative autocorrelation in the sample.*   * 2 is no autocorrelation. * 0 to <2 is positive autocorrelation (common in time series data). * >2 to 4 is negative autocorrelation   Durbin Watson statistic value for model2 is **2.221103**, which says there is no auto correlation exist among predictors observations. |  |

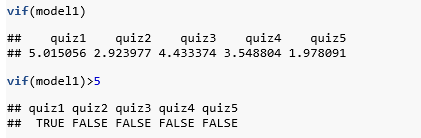


1. What is VIF for each predictor/s? How do you interpret VIF or what VIF signifies? Max 5 lines. [hint: VIF (Variance Inflation Factor)

Show R Output.

**Analysis Response:**

The vif value for model1 is represented in the R output as –



From above variance inflation factor values, variable quiz1 is having more than 5 and hence does represent a factor which influences the R Squared value artificially. Alternately, all the variables other than quiz1 have vif value less than 5 representing no multicollinearity between them.

1. How do you interpret the significance of slope of predictors based on sig. Value or p-value associated with *t*-statistics of each predictor/s. [hint: testing of slope]

Maximum 4 lines. Show R Output.

**Analysis Response:**

The summary display of linear regression models yielded **p-value** for each of the predictors included in the model as part of its output. These p-values are associated with hypothesis testing of predictor variables using t-test method for verification of significant slopes against the response variable.

By default the t-tests are performed at confidence interval 95% (LOS = 5%) with –

|  |  |
| --- | --- |
| H0: | There is no significant slopes for predictor variable against the response variable (*final* in this context) |
| H | There is a significant slope for predictor variable against the response variable (*final* in this context). |

All the predictor variables where the p-value is higher than level of significance, the H0 is accepted indicating that there is no significant slope. However, all the cases where the H0 is rejected, the results can be interpreted as –

*There is a linear relationship between the predictor and the response variable and addition of predictor to the model is likely to be meaningful.*

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response and thus have no significant value add to include in model.

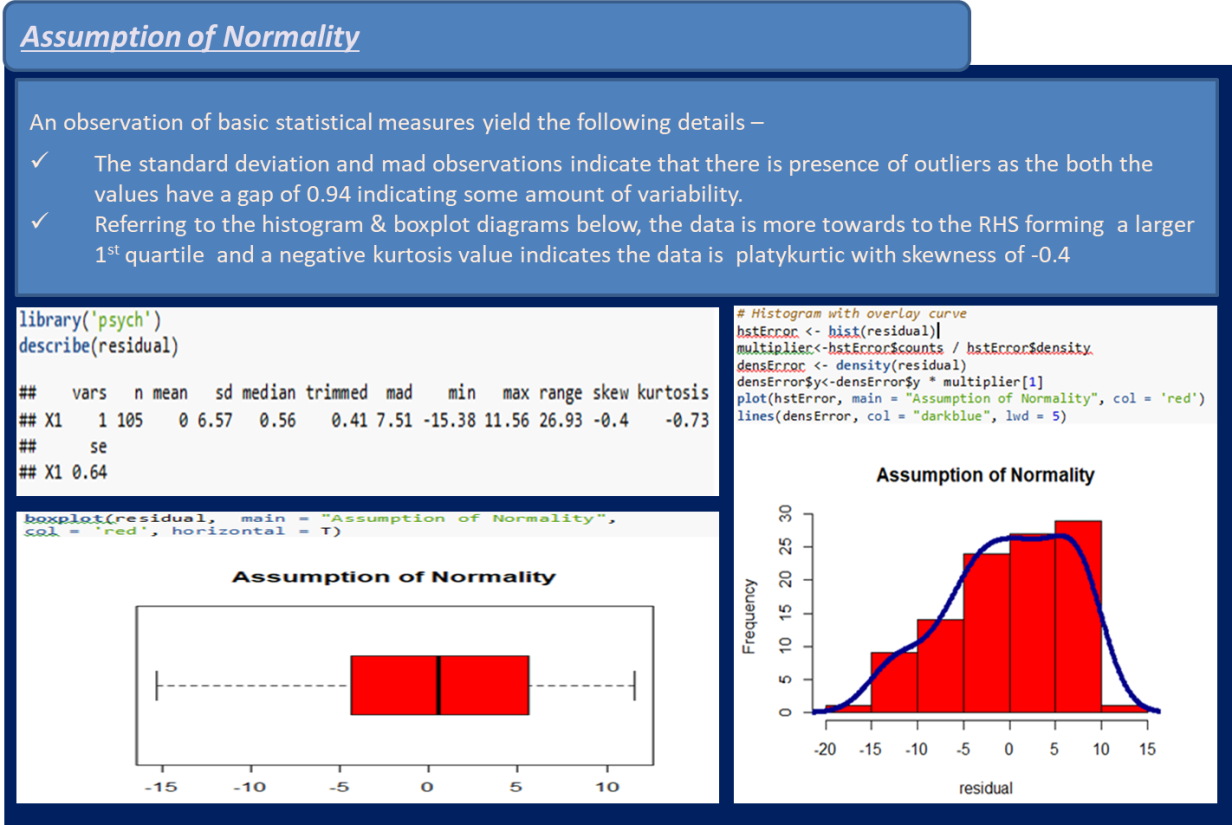
The interpretation for model2 in current context can be expressed as –

|  |  |
| --- | --- |
| * *There is a linear relationship between the predictor variable quiz3 and the response variable final and* * *One unit of change in quiz3 variable results in 1.932 units of change in response variable – final (Since Null hypothesis is rejected only for one variable – quiz3)* |  |

1. Test the assumption of Normality and interpret your findings. [hint: histogram of residuals/errors]

Show histogram and interpret in maximum 3 lines.

**Analysis Response:**



1. Test the assumption of Independent of observations and interpret in maximum 3 lines [hint: draw scatter plot between residuals/errors (y-axis) and observation numbers (x-axis)]

|  |  |
| --- | --- |
| * *The distribution appears unbiased and attains independence* * *The scatterplot doesn’t appear to form any kind of geometry as the points are distributed randomly and freely without any correlation among themselves* |  |

1. Test the assumption of linear relationship and interpret in maximum 3 lines for each predictor [hint: draw scatter plot between response variable, final (y-axis) and predictor/s (x-axis). If more than one predictor is used in model then more scatter plots would be required]

|  |  |
| --- | --- |
| * *The points appear to be symmetrically distributed around a horizontal line in the plot, with a roughly constant variance indicating no violations to linearity.* * *Resemblance of a bowed pattern in the plot diagram indicates that the model makes systematic errors while making large or small predictions* |  |

1. Test the assumption of Constant Error Variance and interpret in maximum 3 lines [hint: draw scatter plot between residuals/errors (y-axis) and predicted values (x-axis)]

|  |  |
| --- | --- |
| * *The points appear to be randomly scattered as function of forecasts/predicted values and are not found to be extreme in both the directions.* * *Errors despite appear to grow larger in one direction – the pattern is not systematic and hence the assumption can be concluded to be in compliance.* |  |

1. What is Standard Error of Estimate of your model and how do you interpret the same. Show with some hypothetical values of predictors. Maximum 300 words. [hint: Standard Error of Estimate]

**Analysis Response:**

The standard error of estimate for the model3 (best fit model) in current context is **6.607** and is used to determine the prediction range using the predicted value obtained from regression equation.

In a regression model, the predictions against a given set of predictors is explained with the help of Adjusted R Squared statistics by performing linear regression analysis over a sample from actual data population, however an important indicator of how precise a prediction about population parameter the sample statistic is, can be measured in terms of **standard error of estimate**.

In other words, the **standard error of the estimate** is a measure of variability of predictions in a regression model that cannot be explained with help of Adjusted R squared values.

Due to the existence of standard error of estimate the predictions derived generally indicate a notional figure and a range is required to express the

forecast with objectives of approximation in prediction / degrading the risk on accuracy of predictions.

|  |  |
| --- | --- |
|  |  |

The regression equation derived from model3 is: Final Score at quiz3 = 46.061+1.931 quiz3 where 46.061 is intercept and 1.931 is regression coefficient.

Thus, if we assume the score of quiz3 to 7 out of 10, the anticipated final score (*Yhat* ) can be calculated as = **46.061+ 1.931 \* 7 = 59.578.**

This prediction of final score cannot be expressed as definitive value due to the effect of standard error of estimate as mentioned above and the range is obtained as follows –

*Upper Limit = Yhat + Critical value of Z at 95% Standard Error or Estimate*

UB = 59.578 + 1.96 \* 6.607 = 72.52772

*Lower Limit = Yhat - Critical value of Z at 95% Standard Error or Estimate*

LB = 59.578 - 1.96 \* 6.607 =46.62828

**Thus at with quiz3 score of 7, the final score is likely to be in the range of 46.62 to 72.52**

1. Congratulation! You have done a marvellous job indeed and build your first predictive model. M just reminding that regression model is somewhere 50% of a data analyst routine job and has great importance in practical world.

Now write a summary of your findings in 250 words which you will show to your reporting manager (before forwarding the model to your client/Principal in this case). This time, no R Output and minimum pictures are needed. Mind it, your reporting manager is a senior statistician/data scientist and do not have time to go into your entire work. He will prefer to read meaningful, to the point and technically correct summary! Here is your chance to impress your boss!

**Analysis Response:**

The given sample of student’s performance records post statistical analysis have yielded some interesting observations and enables to draw some meaningful conclusions about performance pattern of students. The results also enable to discover most affected portions and identify appropriate action points to improve the performance. The results are elaborated as below –

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| * The set of predictors chosen for the linear regression analysis models are – quiz1, quiz2, quiz3, quiz4 and quiz5 for the response variable final at confidence interval of 95% * The best influencer is found to be *quiz3* influencing response variable *final* where 30% of variance can be explained by *quiz3* predictor and rest 70% cannot be explained due to other environmental/external factors outside the model. * Post verification against artificial influencing factors and significant slopes –variable *quiz3* appeared to have significant slope on response variable * While quiz1 has failed the vif test for significant artificial influence whereas quiz2, quiz4 and quiz5 did not possess significant slopes. These finding were also acknowledged by the gap in R squared and adj. R squared values. | **Analysis Records**  Key facts of the statistical analysis performed as part of study can be summarized as below –   |  |  | | --- | --- | | Best Model Derived | (Final ~ quiz3) | | Regression Equation | Y = 46.061+1.931 quiz3 | | Confidence Interval | 95% (LOS = 5%) | | Intercept | 46.061 | | Coefficient of Regression | 1.931 | | Standard Error of Regression | 0.2807 | | Standard Error of Estimate |  | | Adjusted R Squared Value | 0.3083 | |

1. This is final stroke! Besides your boss, your client is equally or rather more important to you!

Your challenge is this that the Principal/client is not statistics savvy! You need to summarize your work/findings in a non-statistical manner or in a lay man manner and this is indeed challenging. However, no way out and you have to do it in a simple but impressive manner (impressive to client!). Write down summary in 500 words.

**Analysis Response:**

The statistical analysis of facilitated sample of student’s performance records is completed successfully and the findings enabled to identify the key contributing factors directly responsible for accelerated improvement in performance records. This implies that having identified the key drivers of impressive performance of students, with minimal efforts the performance **can be raised significantly up to 30%**. In addition, the analysis also helps to reveal few areas of improvements that post remedial treatment can complement the key contributing factors to improve the overall performance dramatically.

The following points present an elaboration of findings and key take-aways that can be translated to specific action items as a proactive measure for improvements.

|  |  |
| --- | --- |
| **FINDINGS** | **RECOMMENDATIONS** |
| * The results indicate that population of passed students is much higher than failed students and a significant number of students have scored between 7 to 10 which is about 70% to 100%. | * Since quiz3 has been found to be most effective, a root cause analysis to investigate the beneficial elements of quiz3 is required to understand the present strength and improvement plans to capitalize the same to be discovered / implemented. |
| * The results indicate that quiz3 has been very significant contributor to student’s present performance and is alone contributing to approx. 30% of positive influence on final score that indeed is contributing to overall performance records. | * Population from section 3 category and gender 1 category to be addressed on high priority (for immediate impact) to discover the key reasons of lagging behind for and should be included in improvement plans identified in above point. |
| * Quiz1 appeared to be contributing to overall performance scores without any significant impact of its own, similarly, quiz2, quiz4 and quiz5 appeared to have nearly no relevance in significant contribution to overall score. | * An analysis is recommended to be performed to discover the key reasons of non-effectiveness of quiz1, quiz2 quiz4 and quiz5 and appropriate measures are to be taken to improve its relevance towards overall score and effectiveness. |
| * Among the population appeared in the exam, population for gender 2 category, population for section 2 category and population for ethnicity 2 category have been more consistent in scoring for quiz3, | * A detailed analysis is recommended to identify the reasons for non-consistency, poor performance of population from section 1 category and ethnicity 1 category and appropriate improvement plans have to be identified |
| * The population from gender 1 category has been effective however has not been found to be consistent for quiz3 |  |
| * Population from section 3 has been approx. close to section 2 in terms of consistency for quiz3. Section 1 and ethnicity 1 however reflected a severe negative growth / inconsistency towards overall performance. |  |

1. Now time to show case your work to rest of the world! Prepare a website as per the sample attached which is only a guideline. Apply your creativity and make it really impressive. This you must attach with your resume in the shape of giving a link in CV. A worth doing exercise.

You may educate your school going wards about individual website and encourage them to show case their projects this way. His/her teacher will be amazed and you will be called by the concerned teacher and head of institution for a thanks giving session! [hint: ppt as guideline and ***Project of Kamana:***[*http://kamanabaproject.wix.com/mmsbaproject*](http://kamanabaproject.wix.com/mmsbaproject)

***Project of Kalyani:***[*http://kalyaninerellaba.wix.com/mmsbaproject*](http://kalyaninerellaba.wix.com/mmsbaproject) Spend some time in viewing the contents of these websites]

**Analysis Response:**

Website link will be shared