**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.

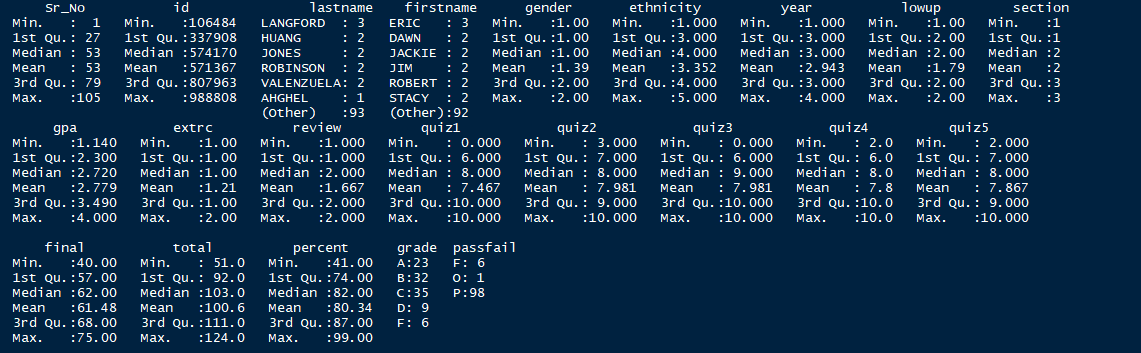
1. Describe data of response variable and predictors in terms of key summary statistics like mean, mode, median, standard deviation, range, skewness and kurtosis. Show histogram and box plots also for each variables. [hint: describe command in R]

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 – |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

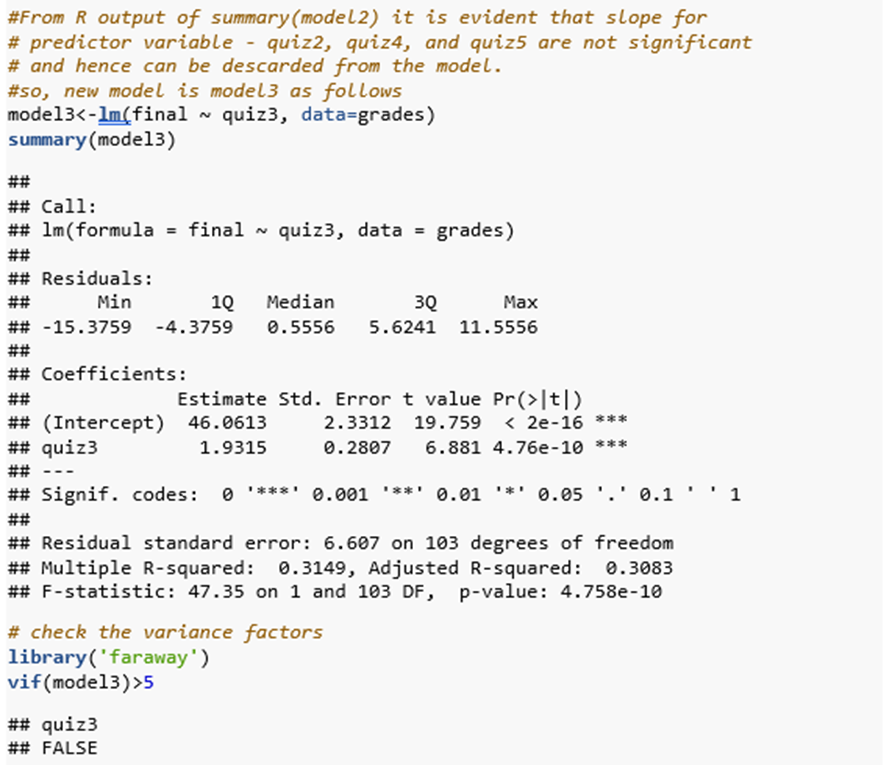
1. What are R Square and Adjusted R Square of your final model? Show R Output and explain in 3 or 4 lines. [hint: R Square and Adjusted R Square]

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 Model** | **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 improve your existing model. That is the desired property of a goodness-of-fit statistic and thus is superior to R Squared.

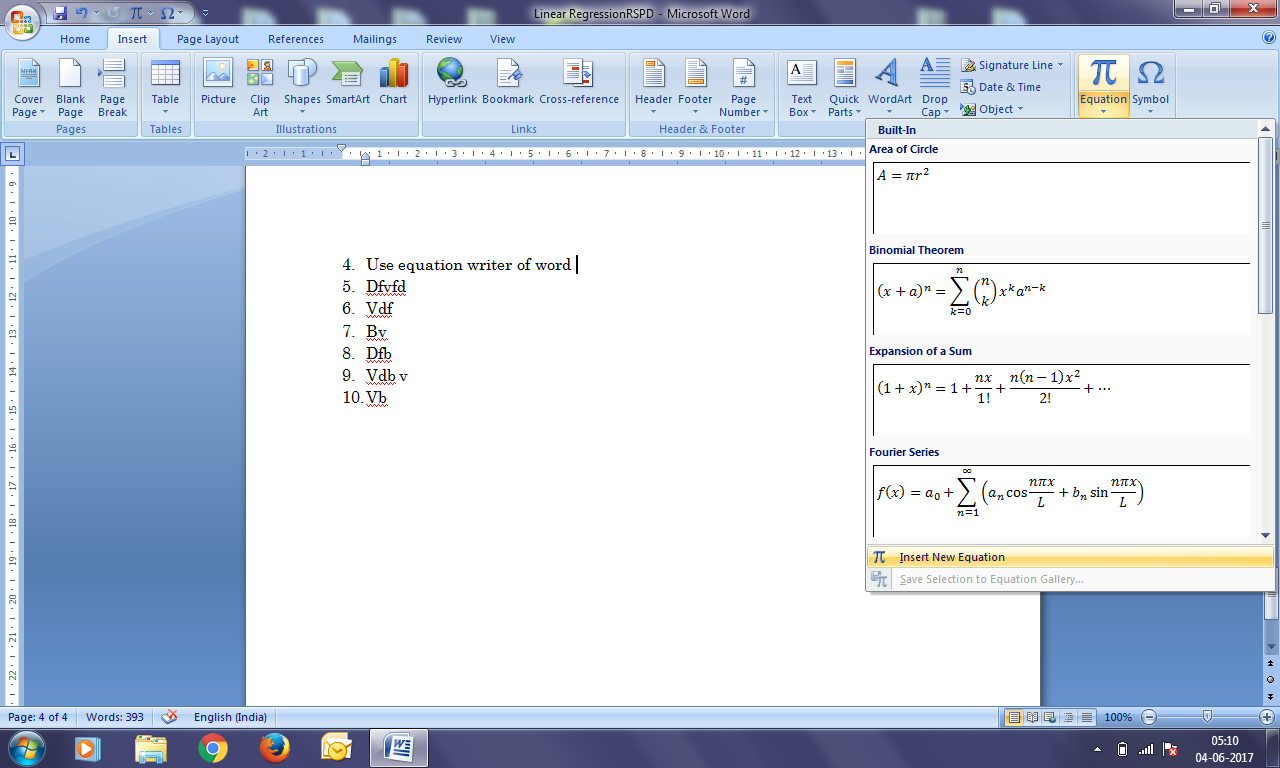
Typically, the more non-significant variables present into the model, the gap in R-squared and Adjusted R-squared is reflected.

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 of the 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 –

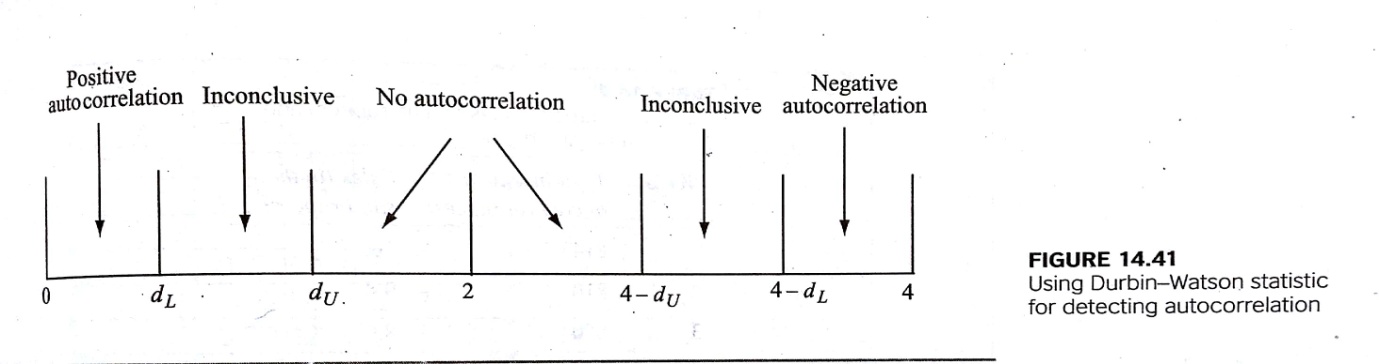
46.061 is the intercept

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:**

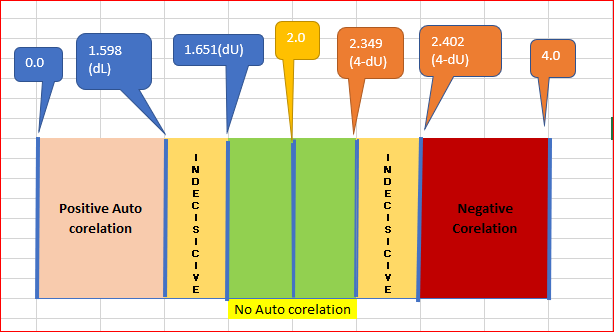
The Durbin Watson statistic is a number that tests for autocorrelation in the residuals from a statistical regression analysis. The Durbin-Watson statistic is always between 0 and 4.

A value of 2 means that there is no autocorrelation in the sample. Durbin Watson statistic value for model2 is 2.1946, which says there is no auto correlation exist among predictors observations.

lag Autocorrelation D-W Statistic p-value

1 -0.09929107 2.194562 0.326

Alternative hypothesis: rho != 0

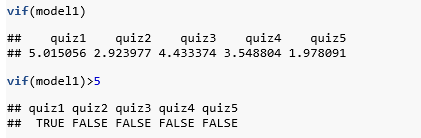


1. What is VIF for each predictor/s? How do you interpret VIF or what VIF signifies? Max 5 lines. [hint: VIF (Variance Inflation Factor)

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:**

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

Show histogram and interpret in maximum 3 lines.

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)]
2. 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]
3. 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)]
4. 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]
5. 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!

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.

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]