**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.

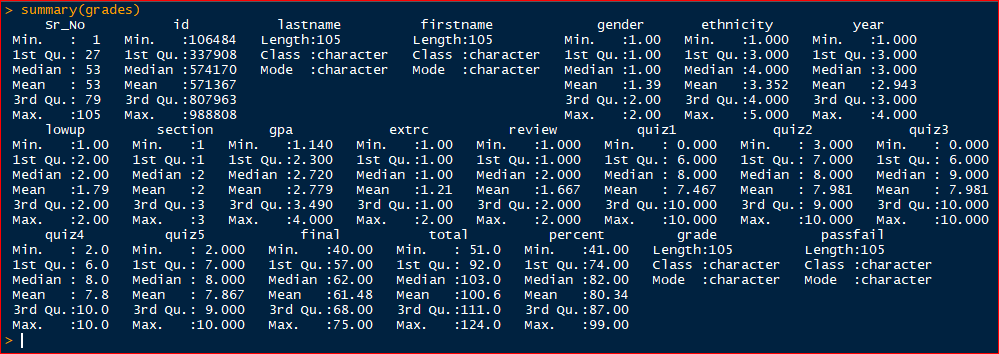
You are required to build at **least 4 models** with different sets of predictors (independent variables). Selection of sets of predictor/s is upon you. Different sets of predictors can be a single variable or more than one variable. However, selection of predictor/s should be based on some logic. For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

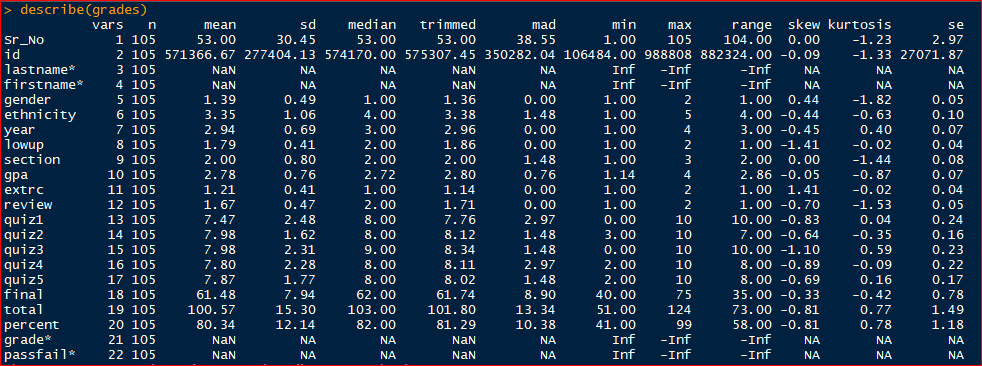
You will analyse all 4 models based on following points and recommend **the best** model to the Principal.

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 variable. [hint: describe command in R]

Each variable to be explained in 30 words maximum.

There are 22 variables in this dataset as shown below





There are 22 variables in data set, 14 variables are nominal data or categorical data, remaining eight variables are scales having meaningful summary and descriptive statistics information.

Eight variables that are having meaningful descriptive statistics are

1.GPA

2.Quiz1

3.Quiz2

4.Quiz3

5.Quiz4

6.Quiz5

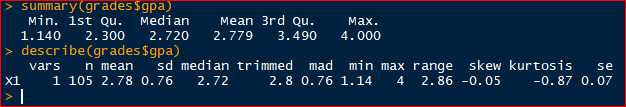
7.Fianl

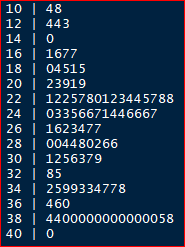
8.Total

Let me explain each one of the above variable using histogram, boxplot, stem & leaf plots with summary and descriptive statistics.

1.GPA:

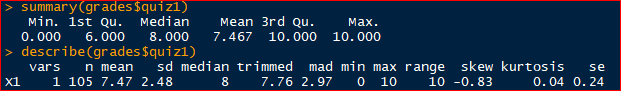
Mean and trimmed mean, SD and mad are very near hence no outliers and no variability, negatively skewed (more data on right side) and data is platykurtic as shown in kurtosis value.

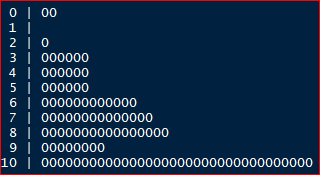




2.Quiz1

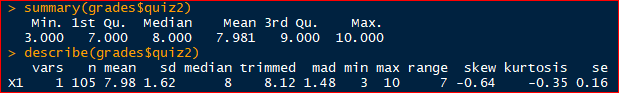
Mean and trimmed mean, sd and mad are very near hence no outliers and no variability, negatively skewed (more data on right side) and data is leptokurtic as shown in kurtosis value

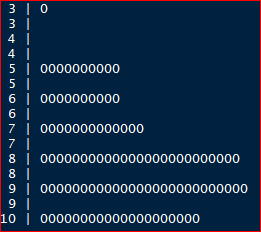




3.Quiz2

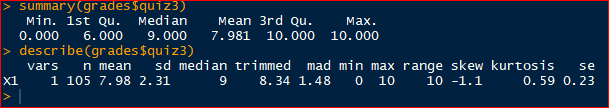
From mean and trimmed mean, sd and mad are differ each other hence there is one outlier, negatively skewed (more data on right side) and data is platykurtic as shown in kurtosis value

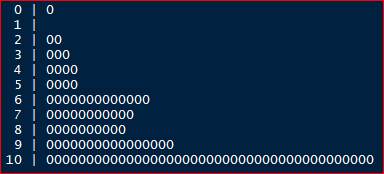




3.Quiz3

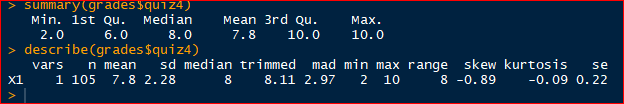
Mean and trimmed mean, sd and mad are very near hence no outliers and no variability, negatively skewed (more data on right side) and data is platykurtic as shown in kurtosis value

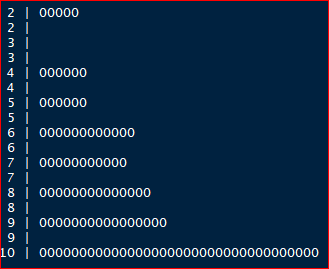




4.Quiz 4

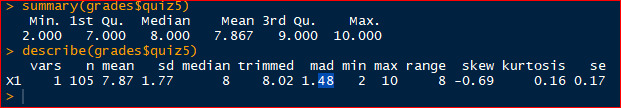
Mean and trimmed mean, sd and mad are very near hence no outliers and no variability, negatively skewed (more data on right side) and data is platykurtic as shown in kurtosis value

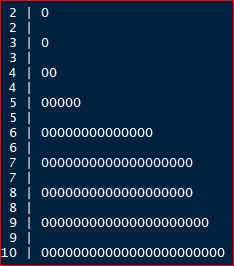


5.Quiz 5

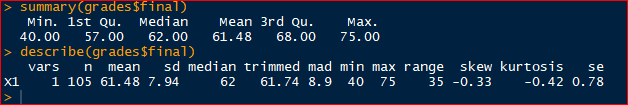
Mean and trimmed mean, sd and mad are different hence two outliers observed negatively skewed (more data on right side) and data is platykurtic as shown in kurtosis value

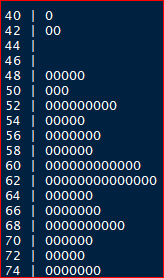




6.final

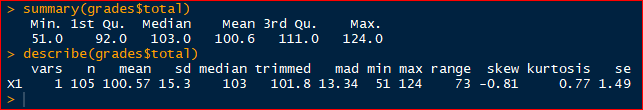
Mean and trimmed mean, sd and mad are different hence one outlier observed most data on right side and data is platykurtic as shown in kurtosis value

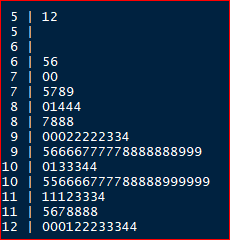


7.Total

Mean and trimmed mean, sd and mad are different hence two outliers observed, negatively skewed (more data on right side) and data is leptokurtic as shown in kurtosis value





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

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

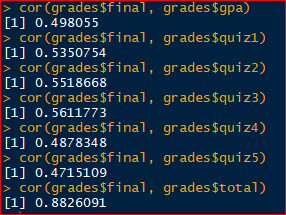
All predictors related positively with response variable(final), variables quiz3 and total are highly correlated with a correlation of 0.5612 and 0.883 respectively.











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.

Final model is model4 and R-squared and Adjusted R-squared values are 0.4535 and 0.4428

Call:

lm(formula = final ~ gpa + quiz3, data = grades)

Coefficients:

(Intercept) gpa quiz3

37.541 3.993 1.609

Call:

lm(formula = final ~ gpa + quiz3, data = grades)

Residuals:

Min 1Q Median 3Q Max

-14.5260 -3.2000 -0.0803 4.7388 12.3955

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 37.5415 2.6802 14.007 < 2e-16 \*\*\*

gpa 3.9926 0.7850 5.086 1.67e-06 \*\*\*

quiz3 1.6088 0.2598 6.193 1.26e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.929 on 102 degrees of freedom

Multiple R-squared: 0.4535, Adjusted R-squared: 0.4428

F-statistic: 42.32 on 2 and 102 DF, p-value: 4.133e-14

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Variable | R-Square | Adj.R-Square |
| 1 | GPA+Quiz1+Quiz2+Quiz3+Quiz4+Quiz5+Total | 0.9735 | 0.9715 |
| 2 | GPA+Quiz1+Quiz2+Quiz3+Quiz4+Quiz5 | 0.4922 | 0.4611 |
| 3 | GPA+Quiz2+Quiz3+Quiz4+Quiz5 | 0.4904 | 0.4647 |
| 4 | GPA+Quiz3 | 0.4535 | 0.4428 |

From the above table high R-Square and Adj.R-Square with a minimum number of variables is model4, final model will be considered after vif and durbin Watson test.

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

One major difference between R Square and the adjusted R-squared is that R-squared supposes that every independent variable in the model explains the variation in the dependent variable, whereas the adjusted R-squared gives the percentage of variation explained by only those independent variables that in reality affect the dependent variable.

R Square is superior due to it is more refined approach to specific variable. R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model

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

F Statistic gives us a power to judge whether that relationship is statistically significant in other words it comments on whether or R² is significant or not.

From anova test, p value it can conclude that quiz2, quiz4 and quiz5 are not having significant slope with response variable final

Response: final

Sum Sq Df F value Pr(>F)

gpa 719.8 1 21.3101 1.174e-05 \*\*\*

quiz2 64.5 1 1.9102 0.17005

quiz3 194.2 1 5.7503 0.01836 \*

quiz4 0.1 1 0.0023 0.96197

quiz5 41.5 1 1.2293 0.27024

Residuals 3343.8 99

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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]

6. 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.

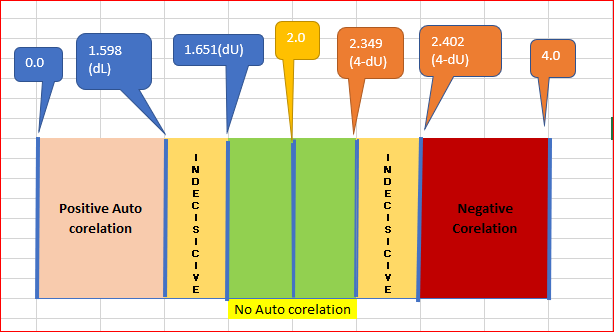
The Durbin Watson statistic is a number that tests for autocorrelation in the residuals from a statistical regression analysis. The Durbin-Watson statistic is always between 0 and 4.

A value of 2 means that there is no autocorrelation in the sample. Durbin Watson statistic value for model3 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



7.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.

VIF of GPA and Quiz3 were shown in R output as follows

gpa quiz3

1.06341 1.06341

From above variance inflation factor values GPA and quiz3 are having less than five hence model3 does not have multicollinearity between predictors as a result no increasing R Squared value artificially, and the VIF values obtained for model3 are as shown above. So, there is no artificial increase in R-Squared value due to correlation among predictors

8. 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.

t-statistics test of multiple regression checks given predictors had significant slope with response variable. If predictors are having p value less than alpha then we can conclude that predictors are having significant slope (there is association) with response variable.

Call:

lm(formula = final ~ gpa + quiz3, data = grades)

Coefficients:

(Intercept) gpa quiz3

37.541 3.993 1.609

Call:

lm(formula = final ~ gpa + quiz3, data = grades)

Residuals:

Min 1Q Median 3Q Max

-14.5260 -3.2000 -0.0803 4.7388 12.3955

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 37.5415 2.6802 14.007 < 2e-16 \*\*\*

gpa 3.9926 0.7850 5.086 1.67e-06 \*\*\*

quiz3 1.6088 0.2598 6.193 1.26e-08 \*\*\*

---

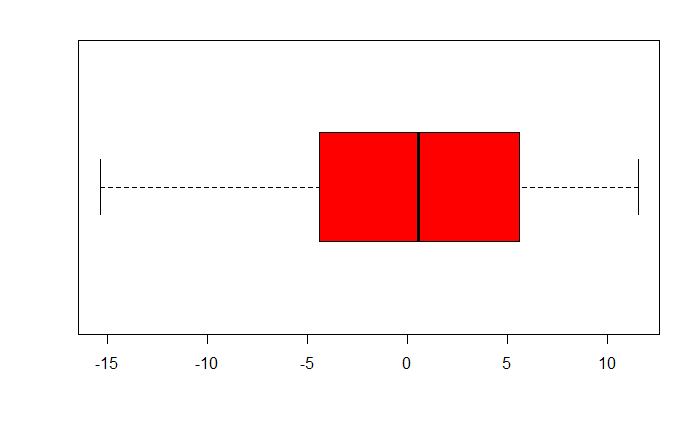
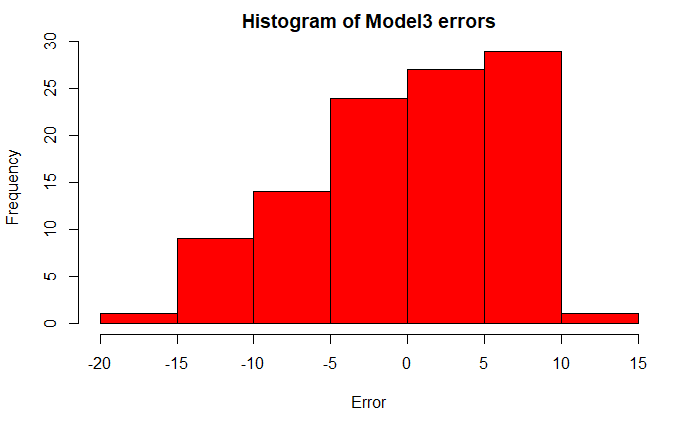
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.929 on 102 degrees of freedom

Multiple R-squared: 0.4535, Adjusted R-squared: 0.4428

F-statistic: 42.32 on 2 and 102 DF, p-value: 4.133e-14

Test the assumption of Normality and interpret your findings. [hint: histogram of residuals/errors]Show histogram and interpret in maximum 3 lines.



9. 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)]

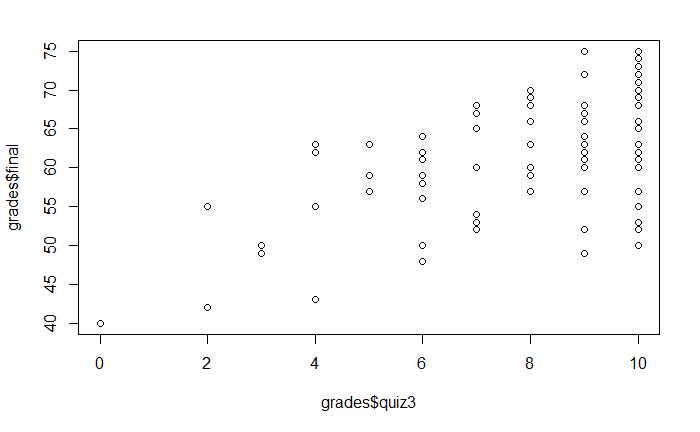
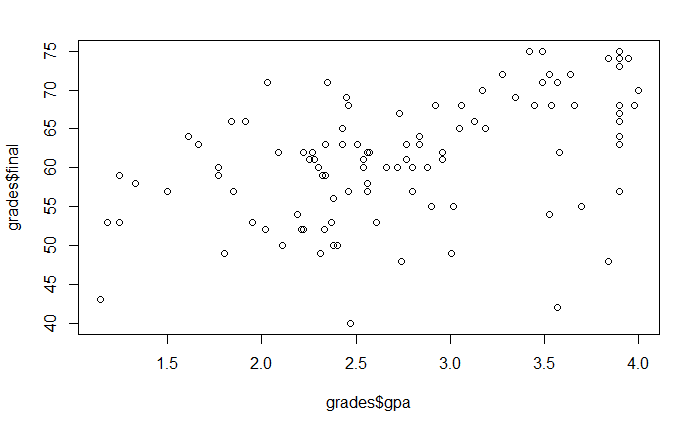
Ans:

Scatter plot of observation number and residual are randomly spread over the graph does not attained any geometry So, assumption of independence of residual assumption passed.

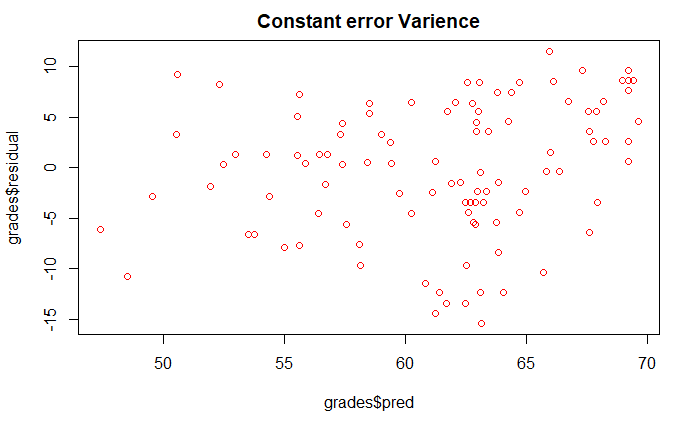


10. Test the assumption of linear relationship and interpret in maximum 3 lines for each predictor [hint: draw scatter plot between response variable, final (y-axis) and predictor/s (x-axis). If more than one predictor is used in model then more scatter plots would be required]

Individually independent variables GPA and Quiz3 correlation found to be insignificant and collectively model built with R-Squared value of



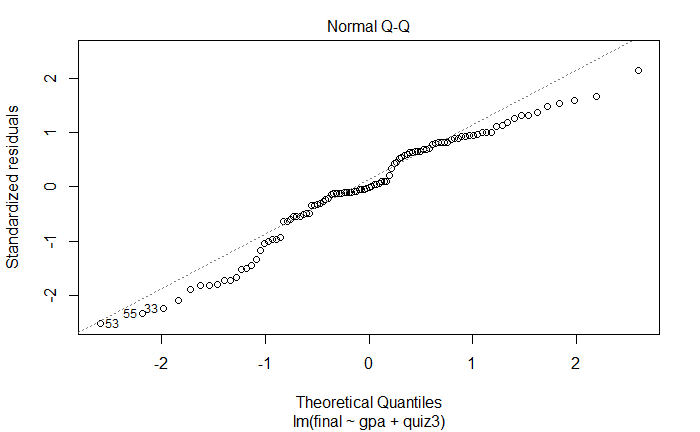
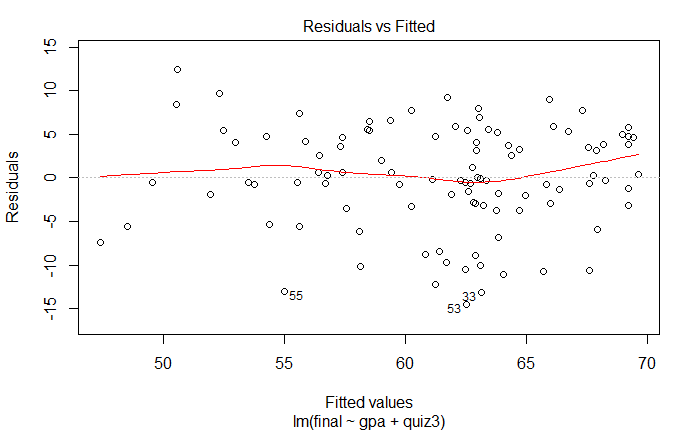
11. Test the assumption of Constant Error Variance and interpret in maximum 3 lines [hint: draw scatter plot between residuals/errors (y-axis) and predicted values (x-axis)]

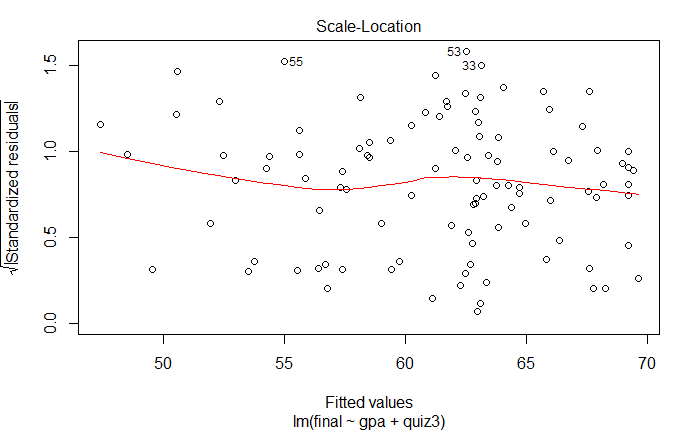
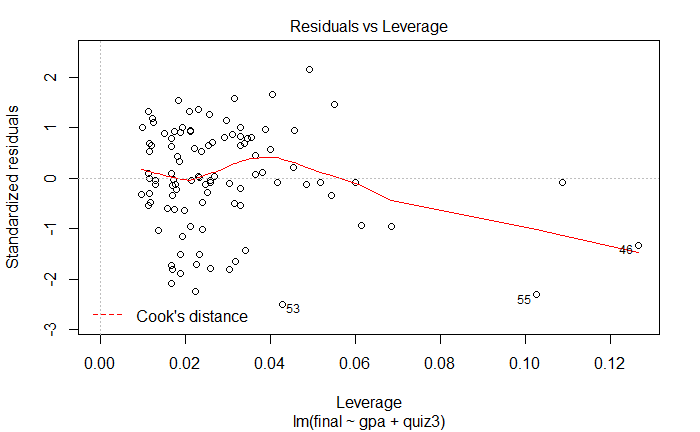


12. What is Standard Error of Estimate of your model and how do you interpret the same. Show with some hypothetical values of predictors. Maximum 300 words. [hint: Standard Error of Estimate]

The standard error of the estimate is a measure of the accuracy of predictions, nothing but sum square error(SSE)In this model considering GPA & quiz3 help us explain 45.35% variance in the final As p value of both is less than LOS (0.05) so we can say that both GPA, quiz3 are significantly The standard error of estimate is a measure of the accuracy of predictions. In a regression line, the smaller the standard error of the estimate is, the more accurate the predictions are. This standard error of estimate is the used to give us the confidence intervals at a particular level of confidence intervals. It is never good to give predictions in the Point estimate form so we give them in a range using standard error of estimate. related to final(dependent variable) .

Following plots are diagnostic plots for our model which describes linear fit is observed.



13. 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!

Ans: Grades.csv consists of 105 observations and 22 variables (mostly categorical or nominal ) .The data tells about the details of the students giving gender, ethnicity, class sections ,marks in 5 quizzes ,marks in final, total marks, gpa and percentage and also tell whether the student is pass or fail .

As we the main motive of the project was to to build a predictive model for predicting final for consumption I have built two linear regression models to predict final

One of our Linear Regression Model predicts the final score using the performance in GPA & quiz3 combined. This multiple regression model helps us explain the variance in the final with 45.35% confidence and the remaining % is due to the other factors like quiz1 quiz2 quiz4 quiz5. The model predictors have significant slope with the dependent or the response variable final which is checked by the significance value for t-test of the predictors i.e the p-value. Moreover the Variance Inflation factor for both the variables is lessthan 5which is in acceptance zone meaning these two predictors are not highly correlated .Also the DWS for the model is 2.19 which lies in the no autocorrelation zone meaning our linear regression is a right model for predicting . Standard error of estimate of our model is 2.6802 means that with this much standard deviation range we would be able to predict our final range with 95% confidence level. The equation of the model of predicting final using GPA & quiz3 with 45.35% variance in final explained is given as below:

The f-test significance value i.e p-value of –f-test is less than LOS meaning that the test of overall model is significant with the final dependent variable.

The assumptions of our models like normality, Linear relationship ,constant error variance are almost satisfied with the data provided and calculated .so we can say that our Model I right way to predict final using quiz2 & quiz3 scores

14. 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.

Model built to predict the final score of students one can predict using GPA & quiz3 marks. Regression equations have been developed and by plugging in the values one can obtain the predicted final score.

Formulae used to predict final score from GPA and Quiz3 derived by the regression model.

Following observtions are made form your school data

* Congratulations principal, your school had highest pass rate at 94.3% which is great achievement to your school.
* In your school no of students of gender-1 and gender-2 has almost same pass rate
* One of the most important observation from the data is students with highest GPA from ethnicity 3 and students GPA above 3.8 are from ethnicity 3 and 4.
* Two students from Ethnicity 3 are(Percentage=99) toppers and more number of students who secured percentage less than 99 above 98 are from ethnicity 4
* In your school more, students are from ethnicity 4 keep admitting ethncity4 and ethnicity 3. It is recommended to give more admission to students from ethnicity 3and 4, as it is seen these students are performed well above 9 in all quiz’s and final too.
* Please look at the following table

|  |  |  |
| --- | --- | --- |
| Grade | No of students | Percentage |
| A | 23 | 21.90 |
| B | 32 | 30.48 |
| C | 35 | 33.33 |
| D | 9 | 8.57 |
| E | 6 | 5.72 |

Coming to performance improvement area’s it recommended to inspire and concentrate to train people from ethnicity 1.

15. 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]