Paresh Patil

Multivariate Linear Regression

* **Reduce Dimensionality**

1. Apply the Missing Value Filter to remove appropriate columns of data.

**Result**: The missing value ratio of “time2” variable was 97.048 % hence we removed that variable from dataset.

1. Apply the Low Variance Filter to remove appropriate columns of data.

**Result**: After applying the Low variance filter I conclude that all the variable has a greater variance and not approximately near to zero.

1. Apply the High Correlation Filter to remove appropriate columns of data.

**Result**: After applying the high correlation filter, I found the correlation of variable “time1” and “time3” is approximately similar and 0.99 near to 1, hence we must remove one variable from dataset, because both are representing same value. Thus, I removed the variable “time3”.

* **Data Transformation**

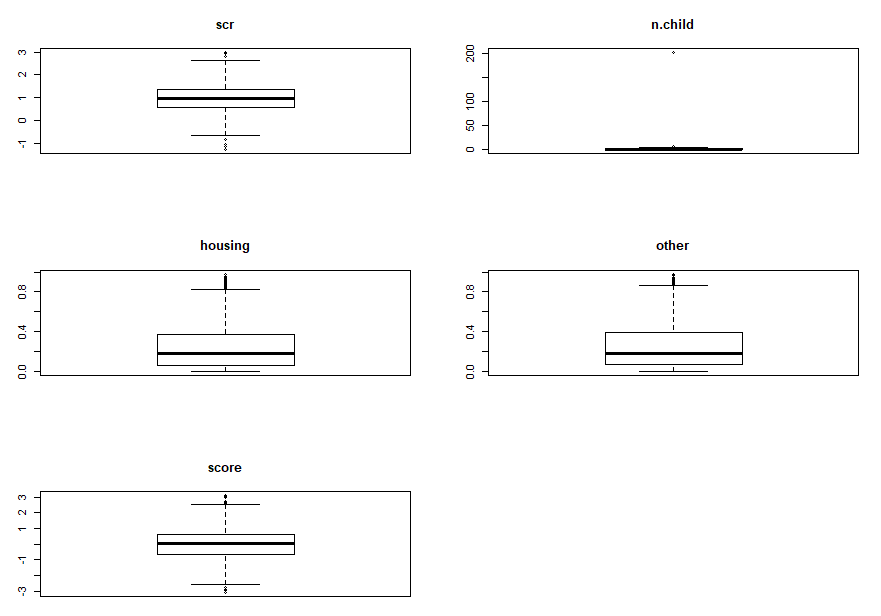
As demonstrated in class, transform any variables that are required to conduct the regression analysis (e.g. categorical variables to dummies).

Result:



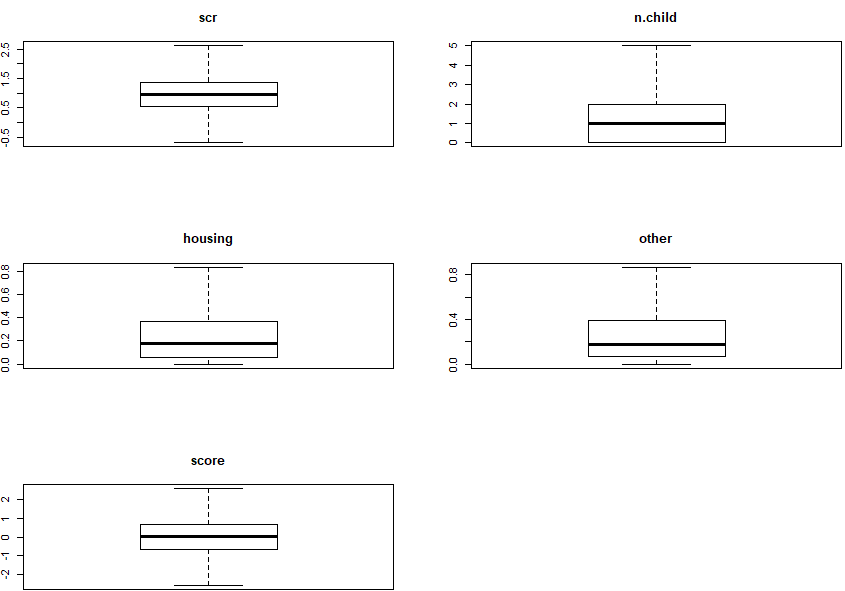
* **Outliers**

1. Create boxplots of all relevant variables (i.e. numeric, non-binary) to determine outliers.



* The “scr” and “score” variable is having a minimal outliers in lower inner and upper inner fence except others.
* “n.child” have a largest outlier in upper outer fence and also in upper inner fence.
* Nonetheless, Outliers of “housing” and “other” lies in upper inner fence.

1. Comment on any outliers you see and deal with them appropriately: Outliers fixed



Here we are fixing the outlier using these formulae

i.e. for lower inner fence

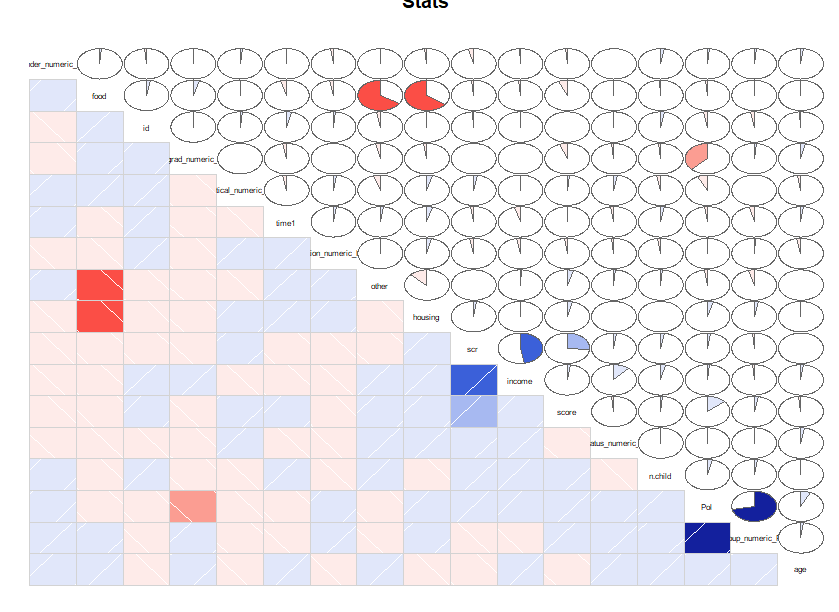
Q3+1.5\*IQR

i.e. for lower inner fence

Q1+1.5\*IQR

* **Exploratory Analysis**

1. Correlations: Create both numeric and graphical correlations (as demonstrated in class) and comment on noteworthy correlations you observe. Are these surprising? Do they make sense?



* After reviewing the corrgram graph, I found very strong positive correlation between “group” and “Pol” variable.
* Similarly, “income” and scr are having moderate positive correlation.
* Surprisingly, “food”~”housing” and “food” ~”other” are having exactly same results. They are having similar moderate negative correlation.
* An there are some variable where is no correlation “age”~”scr”,”age”~”other”,etc
* **Simple Linear Regression**

1. Create a simple linear regression model using Pol as the dependent variable and score as the independent. Create a scatter plot of the two variables and overlay the regression line.

Results: Call:

lm(formula = Pol ~ score, data = MLR\_Filtered\_data\_pp)

Coefficients:

(Intercept) score

* 1. 0.1924

Summary:

Residuals:

Min 1Q Median 3Q Max

-3.2606 -1.0705 -0.0353 0.9591 3.7364

Coefficients:

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

(Intercept) 1.44714 0.03055 47.367 < 2e-16 \*\*\*

score 0.19245 0.03193 6.026 1.99e-09 \*\*\*

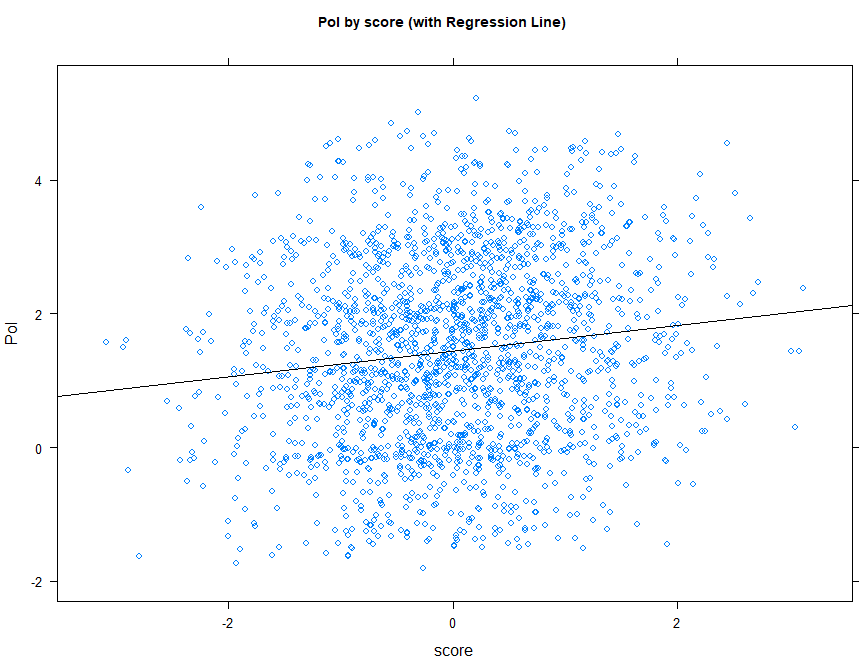
---

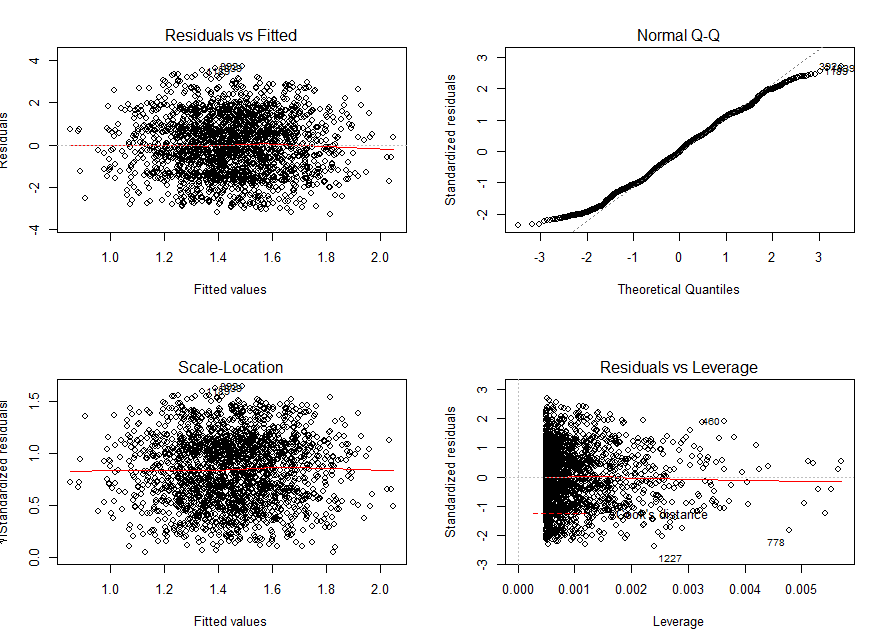
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.377 on 2031 degrees of freedom

Multiple R-squared: 0.01757, Adjusted R-squared: 0.01708

F-statistic: 36.32 on 1 and 2031 DF, p-value: 1.987e-09





1. Create a simple linear regression model using Pol as the dependent variable and scr as the independent. Create a scatter plot of the two variables and overlay the regression line.

Result:

Call:

lm(formula = Pol ~ scr, data = MLR\_Filtered\_data\_pp)

Coefficients:

(Intercept) scr

1.39666 0.05684

Residuals:

Min 1Q Median 3Q Max

-3.2486 -1.0454 -0.0078 0.9955 3.7235

Coefficients:

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

(Intercept) 1.39666 0.05778 24.17 <2e-16 \*\*\*

scr 0.05684 0.05074 1.12 0.263

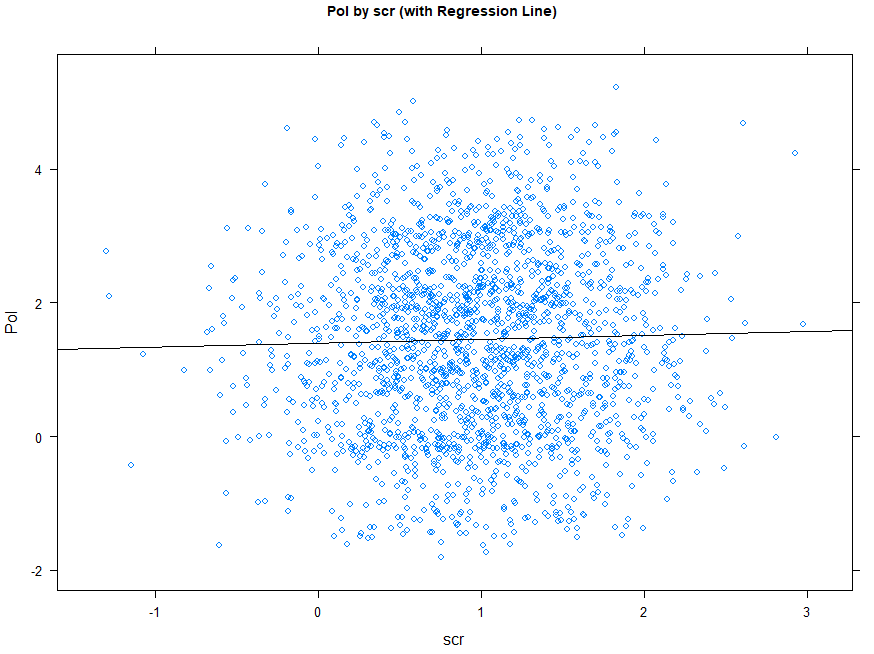
---

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

Residual standard error: 1.389 on 2031 degrees of freedom

Multiple R-squared: 0.0006176, Adjusted R-squared: 0.0001255

F-statistic: 1.255 on 1 and 2031 DF, p-value: 0.2627



1. Compare the models. Which model is superior? Why?

Results: Compare the P values of both the models.

Pol ~scr p-value: 0.2627

Pol~score p-value: 1.987e-09

* Above we clearly see that P value for Linear model of Pol~score is below the significant level, Thus it’s a significant and superior model I choose where as Pol~score is P value is greater than the significant level.
* The reason the model Pol ~scr is not below significant level is the “scr” variable is not significant the p value is greater then the 0.05.
* **Model Development – Multivariate**

1. Backward model:

lm(formula = Pol ~ group + hs.grad + gender + age + political +

income + food + score, data = MLR\_Filtered\_data\_pp, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-0.96028 -0.16694 0.00049 0.17014 1.10628

Coefficients:

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

(Intercept) 8.060e-01 2.745e-02 29.367 < 2e-16 \*\*\*

grouptreat 1.961e+00 1.184e-02 165.628 < 2e-16 \*\*\*

hs.gradyes -1.051e+00 1.187e-02 -88.509 < 2e-16 \*\*\*

gendermale 1.117e+00 1.445e-02 77.276 < 2e-16 \*\*\*

genderundis 6.956e-03 1.454e-02 0.479 0.632

age 5.438e-03 4.301e-04 12.644 < 2e-16 \*\*\*

politicalLiberal -1.194e+00 1.475e-02 -80.966 < 2e-16 \*\*\*

politicalNew\_Democrat 6.637e-03 1.730e-02 0.384 0.701

politicalOther -6.711e-01 1.846e-02 -36.350 < 2e-16 \*\*\*

income 6.493e-07 1.312e-07 4.949 8.09e-07 \*\*\*

food 3.054e-02 2.014e-02 1.516 0.130

score 1.251e-01 6.207e-03 20.149 < 2e-16 \*\*\*

---

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

Residual standard error: 0.2664 on 2021 degrees of freedom

Multiple R-squared: 0.9634, Adjusted R-squared: 0.9632

F-statistic: 4839 on 11 and 2021 DF, p-value: < 2.2e-16

* Residuals: If you observe the min and max residuals of backward model there are no high range difference this indicate that there will no high variation.
* In the result coefficients look at the p value of genderundis, politicalNew\_Democrat , and food are ab0ve the significant level it means there are no influence of this variable in Political Awareness.
* Other all values are below significant level and it show all the variable is having a significant contribution in plotical awareness.
* The deviation of 0.2664 is present rom residual and its mean.
* Our R squared value and Adjust R squared values are almost same. when we are implementing the multiple regression that time, we must consider the Multiple R squared value. When we add a variable either significant or insignificant it will increase the multiple R Squared value, but it will decrease the value of adjusted R squared value. However, our both the values are similar.
* Let’s observe the F stat P value it is below the significant level hence the model is significant.

Note: Above explanation is for both the model “Full” and “Backward” because I could not found any difference both are giving the same results.

1. Full Model:

Residuals:

Min 1Q Median 3Q Max

-0.96028 -0.16694 0.00049 0.17014 1.10628

Coefficients:

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

(Intercept) 8.060e-01 2.745e-02 29.367 < 2e-16 \*\*\*

grouptreat 1.961e+00 1.184e-02 165.628 < 2e-16 \*\*\*

hs.gradyes -1.051e+00 1.187e-02 -88.509 < 2e-16 \*\*\*

gendermale 1.117e+00 1.445e-02 77.276 < 2e-16 \*\*\*

genderundis 6.956e-03 1.454e-02 0.479 0.632

age 5.438e-03 4.301e-04 12.644 < 2e-16 \*\*\*

politicalLiberal -1.194e+00 1.475e-02 -80.966 < 2e-16 \*\*\*

politicalNew\_Democrat 6.637e-03 1.730e-02 0.384 0.701

politicalOther -6.711e-01 1.846e-02 -36.350 < 2e-16 \*\*\*

income 6.493e-07 1.312e-07 4.949 8.09e-07 \*\*\*

food 3.054e-02 2.014e-02 1.516 0.130

score 1.251e-01 6.207e-03 20.149 < 2e-16 \*\*\*

---

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

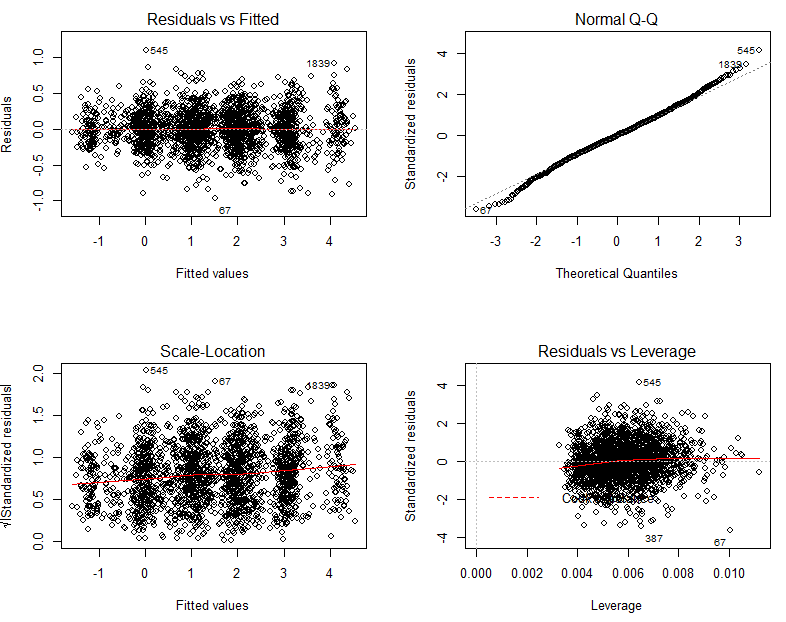
Residual standard error: 0.2664 on 2021 degrees of freedom

Multiple R-squared: 0.9634, Adjusted R-squared: 0.9632

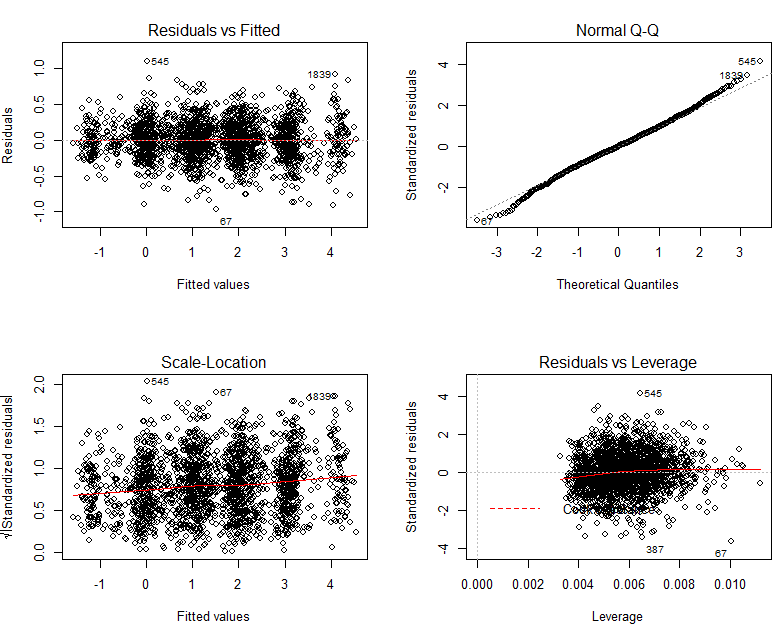
F-statistic: 4839 on 11 and 2021 DF, p-value: < 2.2e-16.

**Model Evaluation – Verifying Assumptions – Multivariate**

1. Backward model:



1. Full Model:



Both the models are choosing the same variables and giving same results and both the model are significant . We can choose any model from above.