Stat 486 Final Report

Xiaohan Liu

Dec 14th, 2017

1. Introduction

The study was based on a two compiled datasets from Statistical Abstract of the United States, 1981, U.S. Bureau of the Census, Washington, D.C. It provided some basic sociological information in 26 metrics for each of the 50 states, and they are listed in Table 1. Two major concerns were raised by our respected clients: 1. Find an appropriate linear regression model with M as the dependent variable along with MA, D, S, B, HT, UR, CR, HS, INC, PL and VT to choose from as independent variables; 2. Find an appropriate linear regression model with MA as the dependent variable along with all other variables to choose from as independent variables.

Statistic	N	Mean	St. Dev.	Min	Max
POP	50	4,511.40	4,723.91	400	23,669
UR	50	669.30	144.01	338	913
MV	50	480.88	65.70	363	731
BL	50	520.78	623.19	1	2,402
SP	50	291.76	784.65	3	4,544
AI	50	283.46	428.73	10	2,013
IN	50	42.42	47.65	1	218
PR	50	65.26	84.14	3	499
MH	50	40.88	46.60	0	223
B	50	168.02	30.36	122	301
HT	50	314.26	69.31	76	431
S	50	130.10	30.81	72	248
DI	50	151.82	36.26	27	252
MA	50	135.50	194.50	75	1,474
D	50	57.78	22.63	30	168
DR	50	161.86	40.15	102	261
DN	50	51.74	10.82	32	74
HS	50	672.92	72.67	523	802
CR	50	5,490.10	1,402.92	2,552	8,854
M	50	81.78	45.13	7	200
PI	50	119.46	55.67	28	244
RP	50	521.10	72.69	372	728
VT	50	551.40	75.14	407	704
PH	50	55.52	5.94	36	66
INC	50	5,129.78	719.77	3,677	7,141
PL	50	115.68	42.19	67	261

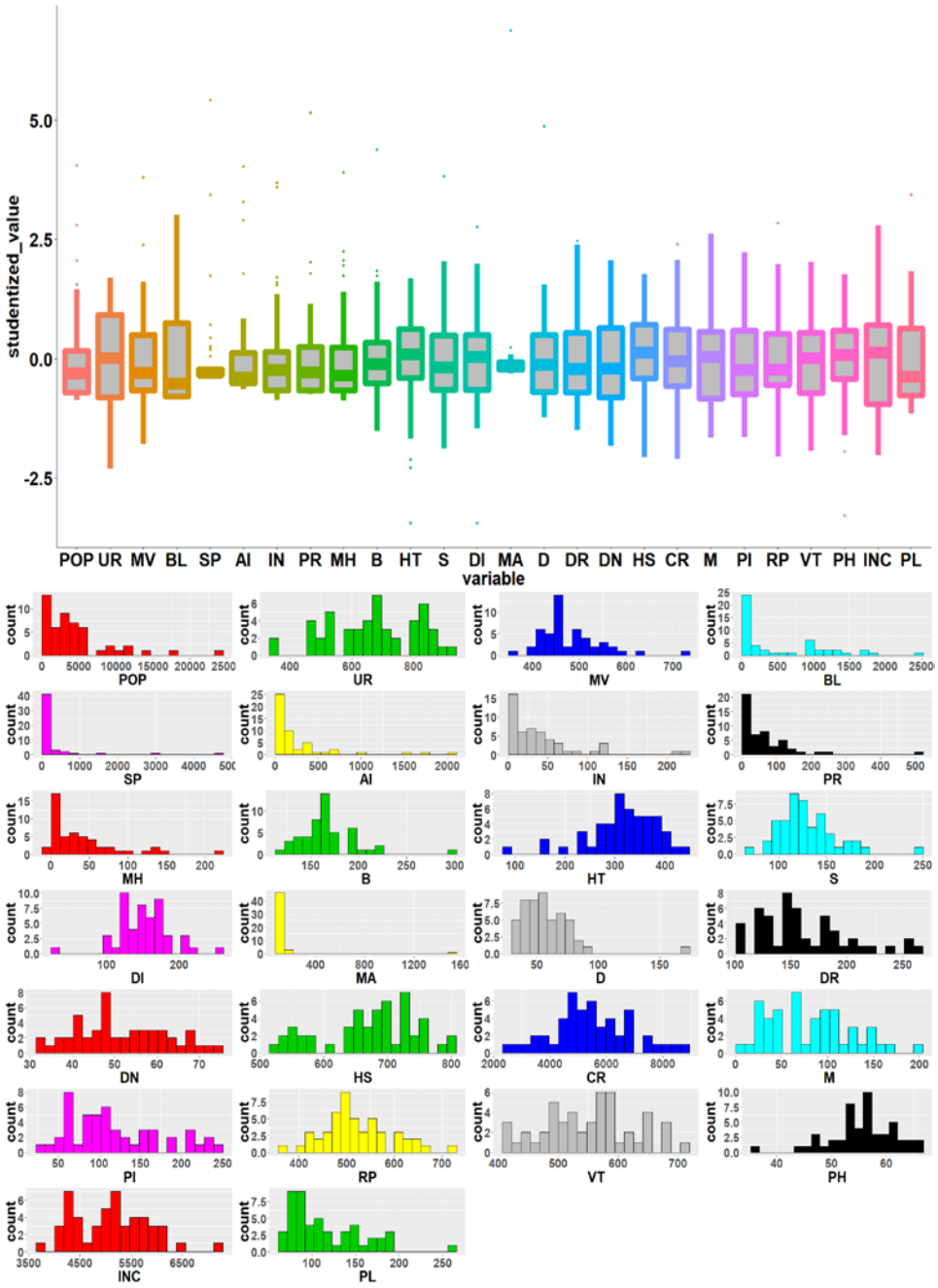
Table 1. Summary statistics on all variables

2. Methodology

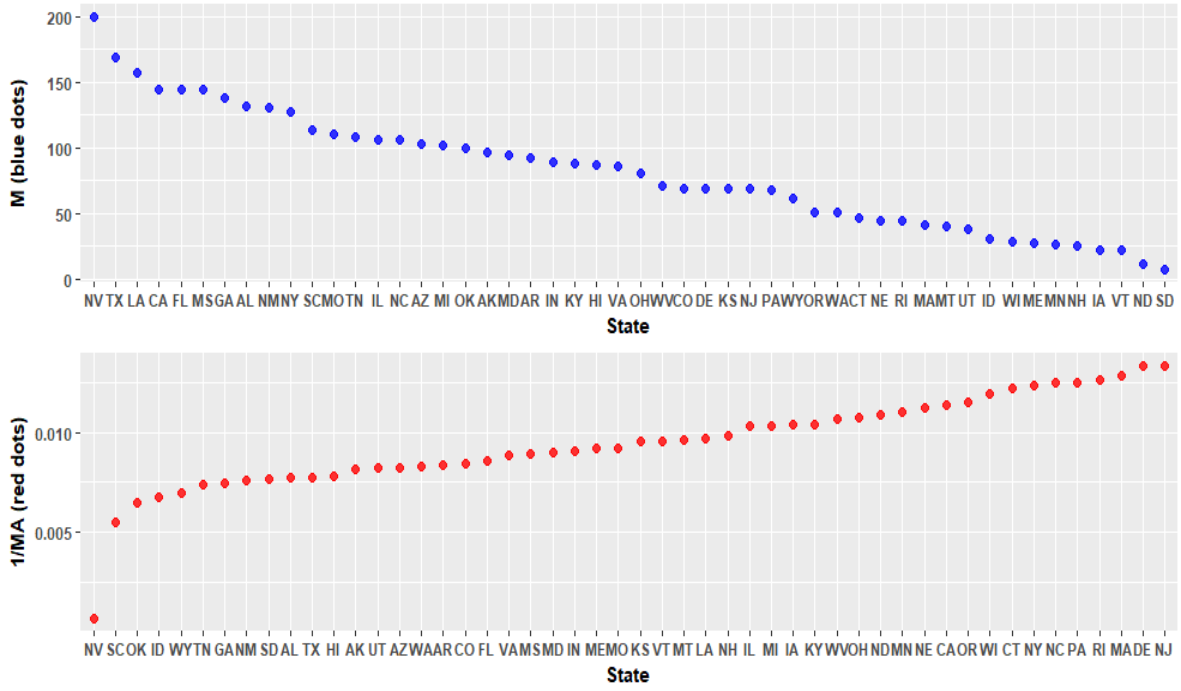
All of our statistical analysis was conducted on the platform R 3.4.3. The procedures were similar between the two topics with the primary difference lied in the transformation.

2.1. Exploratory Data Analysis:

EDA techniques were used to explore the data quality at the first glimpse. Summary statistics of all variables could be found in Table 1, with their graphical displays shown in Graph 1 and Graph 2.



Graph 1. Box plots of all studentized variables and Histograms of all variables



Graph 2. Scatter plot on M and 1/MA from each state

2.2 Test of Model adequacy:

Starting from this step, we applied standardization on each candidate independent variable respectively to ease further computations and it should change our regression results. In order to know if some transformation on the independent variables and/or dependent variable is necessary, we were using the Near-neighbor approach to test our current model adequacy. The test statistic T for it is:

$$T = \frac{\Omega - E(\Omega)}{Var(\Omega)} \sim t(n-p-1),\ and\ \ \Omega = \frac{e'Ve}{s^2}$$

where,

$$V_{il} = V_{li} = -2\omega_{il};\ and\ \ V_{ii} = \sum_{l} V_{il}$$

$$\omega_{il} = \omega_{li} = \frac{1}{(||x_i - x_l||)^2},\ and\ \omega_{ii} = 0\ \ \forall i, l \in \{1, \cdots, p+1\}$$

We had also found that:

$$E(\Omega) = tr(B);\ \ and\ Var(\Omega) = \frac{2[(n-p-1)tr(B'B) - (tr(B))^2]}{n-1}$$

where,

$$B = M'VM;\ and\ \ M = I - H_X$$

And, we would reject the model adequacy if $|T| > t_{\alpha/2,n-p-1},\ \alpha = 0.05$.

2.3. Transformation

If the current model fails the above adequacy test, we would do some transformation, such as the derivatives of BoxCox power transformation, on its response first and then on the independent variables if necessary until the transformed model passed the near-neighbor approach test for adequacy.

2.4. Step-wise variable selection

After we obtained an adequate model, we would perform dimension reduction by a step-wise both-direction method with p-values and AIC as our primary criteria along with adj-R^2 and C(p) as secondary criteria.

2.5. Multicollinearity diagnostics

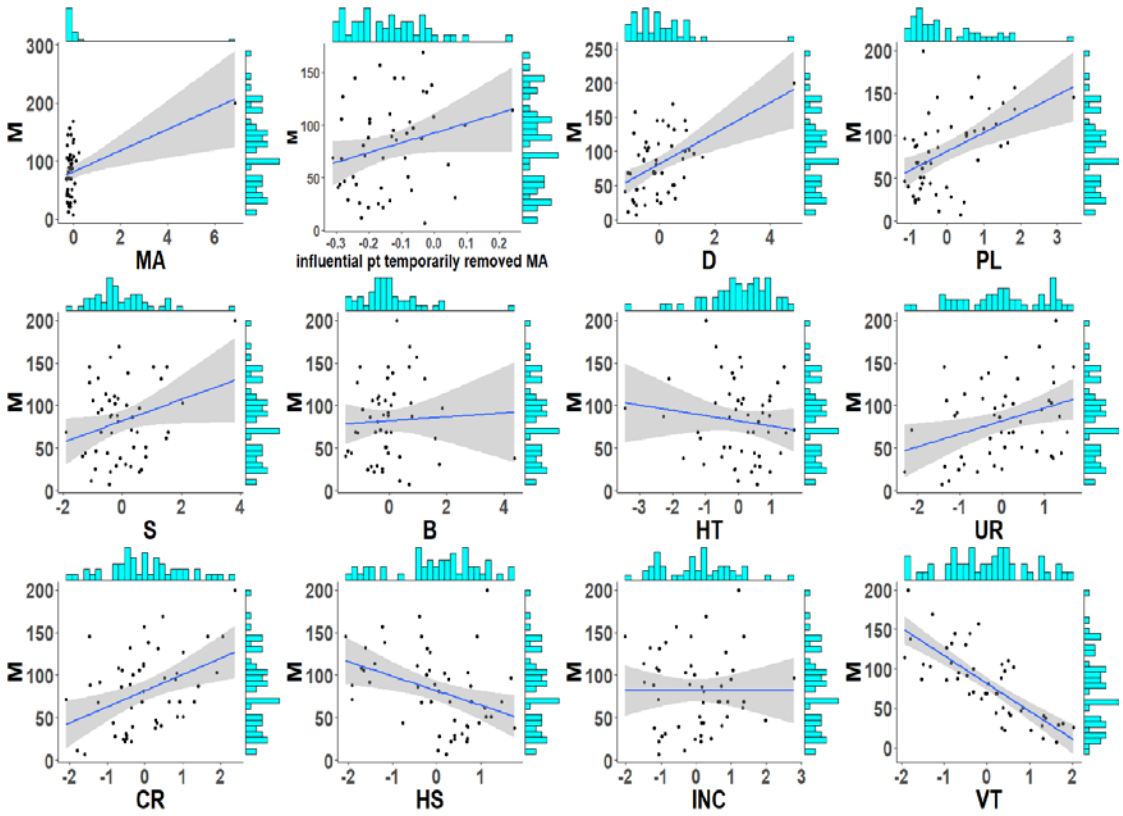
Upon the model with selected variables, we would utilize VIF (Variance Inflation Factor) and Condition Index side by side to detect the variable(s) influenced by multicollinearity. Our criteria were: From variables whose VIF are large than 10 or Condition Index are larger than 40, we would delete the variable with the largest VIF and Condition Index until none exists. Here, we choose to simply delete variable is given our number of observation is relatively small compared to the number of independent variables.

2.6. Linear regression, residual diagnostics and influence measures

Based on the eventually screened out independent variable, we would do a linear regression with all interaction terms automatically into consideration. Then we would test the linearity by the residual vs. fitted values plots, the normality by qqplot along with Shapiro-Wilk test and Kolmogorov-Smirnov test. Besides, we would carry out the measures of influence via Cook's D plot, studentized residual plot, deleted studentized residual vs. fitted values plot and studentized residuals vs leverage plot to detect the influential points and outliers.

3. Results and Discussion

3.1. Find an appropriate linear regression model with M as the dependent variable along with MA, D, S, B, HT, UR, CR, HS, INC, PL and VT to choose from as independent variables.



Graph 3. Partial scatter plots in 1st problem

In the initial EDA, we checked the partial scatter plots of each candidate independent variable versus the responses like in Graph 3. Aside from one influential point in MA screwing up the range of the entire plot, all other plots look decent enough, hinting no need for transformation. And it was checked by another partial plot between influential point temporarily removed MA and the response M, and the same could be expected for variable D. We then carried out the model adequacy test by the Near-neighbor approach, and got a well above any common significance level p-value of 0.50. This was in agreement with our previous guess that no transformation was required on the as-is linear model.

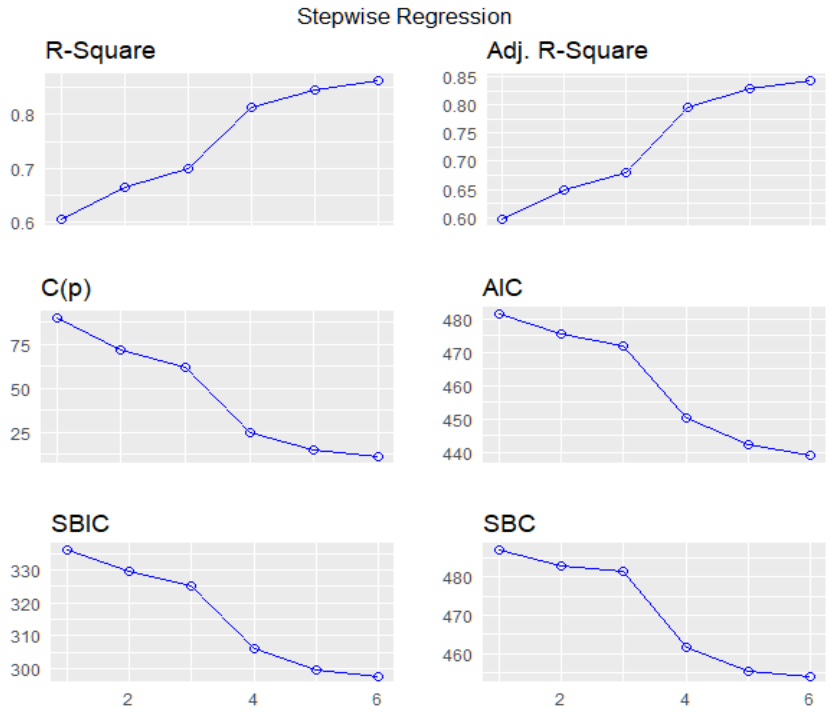
Next, we proceeded with the variable selection from all candidates following a step-wise both-direction fashion evaluated by various different criteria. The results are listed below in Table 2. :

```
## Stepwise Selection Method

## Candidate Terms:

## 1 . MA
## 2 . D
## 3 . PL
## 4 . S
## 5 . B
## 6 . HT
## 7 . UR
```

```
## 8 . CR
## 9 . HS
## 10 . INC
## 11 . VT

## ----------------------------------------------------------------------------------------
##                                Stepwise Selection Summary
## ----------------------------------------------------------------------------------------
##                      Added/                   Adj.
## Step    Variable     Removed     R-Square     R-Square     C(p)        AIC         RMSE
## ----------------------------------------------------------------------------------------
##    1       VT       addition       0.605       0.596    90.1920    481.4401     28.67
##    2       D        addition       0.663       0.649    72.0210    475.4251     26.74
##    3       PL       addition       0.699       0.679    61.8560    471.8868     25.57
##    4       UR       addition       0.812       0.795    24.7290    450.2473     20.41
##    5      INC       addition       0.847       0.829    14.8760    442.1337     18.65
##    6       HS       addition       0.862       0.843    11.5680    438.8443     17.90
## ----------------------------------------
```

Table 2. Step-wise variable selection by p-value for 1st problem

We also plotted the above variable selection result in Graph 4.



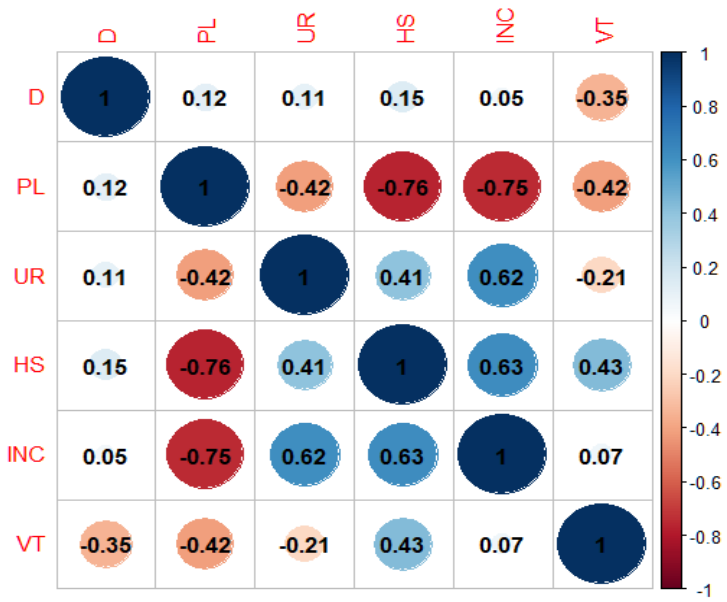
Graph 4. Stewise regression plots for 1st problem

As indicated in methodology, we followed p-values and AIC as our primary criteria along with adj-R^2 and C(p) as secondary criteria. Meanwhile, we were fully aware of that step-wise regression could miss the "best" model. Thus,after consulted with the best subset model selected from all 2^{11} possible models (shown in Table 3. and Graph 5) and the real life meaning of the variables, the final model we landed on was:

$$M \sim D + PL + UR + HS + INC + VT + \epsilon$$

```
##            Best Subsets Regression
## --------------------------------------------
## Model Index    Predictors
## --------------------------------------------
##      1         VT
##      2         D VT
##      3         PL UR VT
##      4         D PL UR VT
##      5         D PL UR INC VT
##      6         D PL UR HS INC VT
##      7         MA PL B CR HS INC VT
##      8         MA PL B UR CR HS INC VT
##      9         MA D PL B UR CR HS INC VT
##     10         MA D PL B HT UR CR HS INC VT
##     11         MA D PL S B HT UR CR HS INC VT
## --------------------------------------------




## Subsets Regression Summary
## ------------------------------------------------------------------------------------
##                        Adj.        Pred

## Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC
------------------------------------------------------------------------------------
##   1        0.6047      0.5964      0.5712    90.1925    481.4401    336.2137
##   2        0.6632      0.6489      0.6319    72.0208    475.4251    329.4694
##   3        0.7552      0.7392      0.7056    42.3469    461.4836    316.0004
##   4        0.8121      0.7954      0.7595    24.7292    450.2473    305.9744
##   5        0.8465      0.8291      0.7963    14.8756    442.1337    299.5814
##   6        0.8619      0.8427      0.8019    11.5677    438.8443    297.6611
##   7        0.8751      0.8542      0.7103     9.0407    435.8440    296.4849
##   8        0.8810      0.8577      0.8016     9.0120    435.4299    297.3848
##   9        0.8858      0.8601      0.6518     9.3463    435.3567    298.7316
##  10        0.8879      0.8592      0.7053    10.6059    436.4069    300.8319
##  11        0.8897      0.8578      0.6412    12.0000    437.6160    303.1016
## ------------------------------------------------------------------------------------
##  AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
```

Table 3. Best of all subsets variable selection for 1st problem



Graph 5. Correlation matrix of the chose model for 1st problem

In addition, we tested the chosen model's multicollinearity in Table 4. All variables' VIFs were well below 10 and condition indices were well below 30. Thus, we could say this chosen model didn't possess a prominent multicollinearity issue.

```
## Tolerance and Variance Inflation Factor
## ---------------------------------------
## # A tibble: 6 x 3
##   Variables Tolerance      VIF
##       <chr>     <dbl>    <dbl>
## 1         D 0.6732573 1.485316
## 2        PL 0.2356492 4.243596
## 3        UR 0.5195508 1.924740
## 4        HS 0.2815639 3.551592
## 5       INC 0.3018392 3.313022
## 6        VT 0.4552208 2.196736
##
##
## Eigenvalue and Condition Index
## ------------------------------
##   Eigenvalue Condition Index intercept            D           PL
## 1  2.8969982        1.000000         0 1.416899e-06 2.355829e-02
## 2  1.5316353        1.375297         0 1.370817e-01 3.333205e-03
## 3  1.0000000        1.702057         1 0.000000e+00 0.000000e+00
## 4  0.8677503        1.827161         0 4.493292e-01 1.465264e-05
## 5  0.3555144        2.854602         0 8.752961e-04 7.554811e-02
## 6  0.1999796        3.806109         0 2.303512e-01 4.778421e-02
## 7  0.1481223        4.422460         0 1.823612e-01 8.497615e-01
```

```
##           UR           HS          INC          VT
## 1 0.02485891 0.0257043229 0.027286286 0.007134506
## 2 0.06108475 0.0006583844 0.006574516 0.133584722
## 3 0.00000000 0.0000000000 0.000000000 0.000000000
## 4 0.10246131 0.0427161956 0.011415125 0.044705752
## 5 0.68945943 0.0268583361 0.201430222 0.182619683
## 6 0.03587572 0.4929433727 0.378436527 0.629130843
## 7 0.08625988 0.4111193883 0.374857324 0.002824495
```

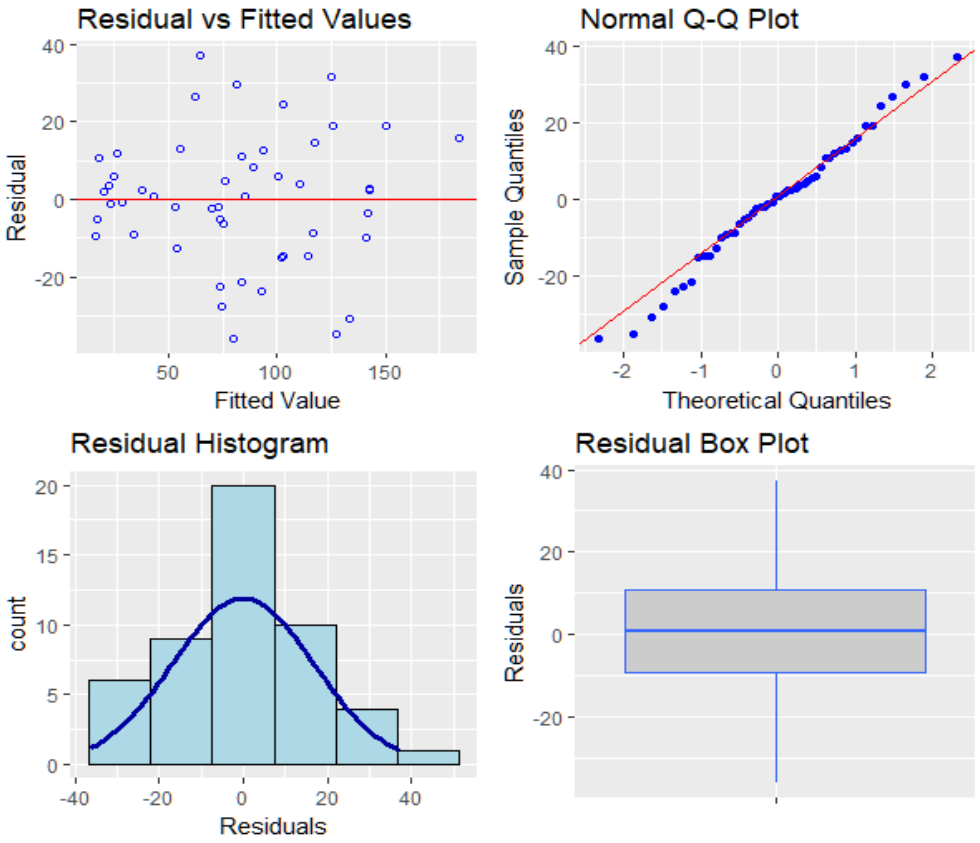
Table 4. Multicollinearity test for 1st problem

Then, the chosen model's linear regression result was displayed in Table 5, including the model summary, ANOVA table and the parameter estimates including the 95% C.I.. It should be noted although we included all the interaction terms but none was significant, meaning all the independent variables tended to influence the response independently. The adjusted R^2 value was 0.843, demonstrating a high level goodness-of-fit. From the residual diagnostic (Graph 6.), we could observe no clear pattern, indicating the validity of the linearity. The normality assumption was also supported by the qqplot and histogram, and further corroborated by S-W test (p-value:0.919) and K-S test (p-value:0.914). Meanwhile, form the influence measures (Graph 7.), we found observation 33 (NV) stray furthest from the herd in both Cook's D and Leverage metric. Under closer examination, it was mainly due to its absurdly high divorce rate (studentized value equals 4.87), which could be justified probably only by the very existence of Las Vegas and its one-of-a-kind social ecological model. Furthermore, DFBETAs (Graph 8.) measured the difference in each coefficient estimate with and without the influential point. From which, we could get senses on how much one observation has effected the estimate of a regression coefficient. Clearly, apart from the NV, observation 39 (RI) also looked have huge influences on variable PL, UR, HS, INC and VT. On the other hand, we should expect 2.5 outliers on average out of our 50 observations. In practice, 4 were found (Graph 7.), and our dear clients should treat them with extra care. In particular, the "worst" one of them, possessing both the largest leverage and studentized residual was observation 39 (RI), a.k.a the tiniest state of all, was not extremely surprising. Hence, if it was necessary to remove this outlier, it might result in one of the slightest possible impacts to the validity of this model nation-wide.

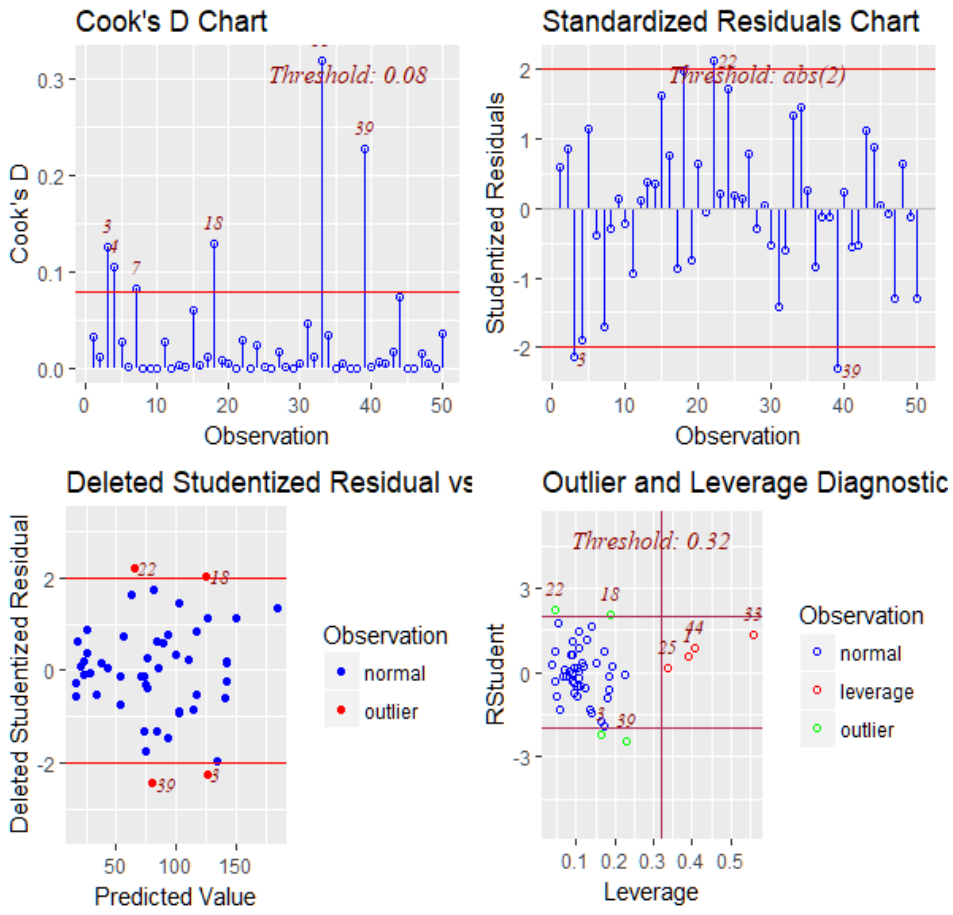
```
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.928       RMSE                17.902
## R-Squared               0.862       Coef. Var           21.891
## Adj. R-Squared          0.843       MSE                320.484
## Pred R-Squared          0.802       MAE                 12.821
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## --------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
```

```
## Regression    86027.766         6      14337.961    44.738    0.0000
## Residual      13780.814        43        320.484
## Total         99808.580        49
## --------------------------------------------------------------------
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model        Beta     Std. Error    Std. Beta       t        Sig        lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)     81.780          2.532                  32.302    0.000      76.674    86.886
##           D     13.996          3.117         0.310     4.491    0.000       7.711    20.282
##          PL     26.048          5.268         0.577     4.944    0.000      15.423    36.672
##          UR     16.208          3.548         0.359     4.568    0.000       9.053    23.363
##          HS    -10.557          4.820        -0.234    -2.190    0.034     -20.277    -0.838
##         INC     16.272          4.655         0.361     3.496    0.001       6.884    25.659
##          VT    -12.552          3.790        -0.278    -3.312    0.002     -20.197    -4.908
## ----------------------------------------------------------------------------------------
```

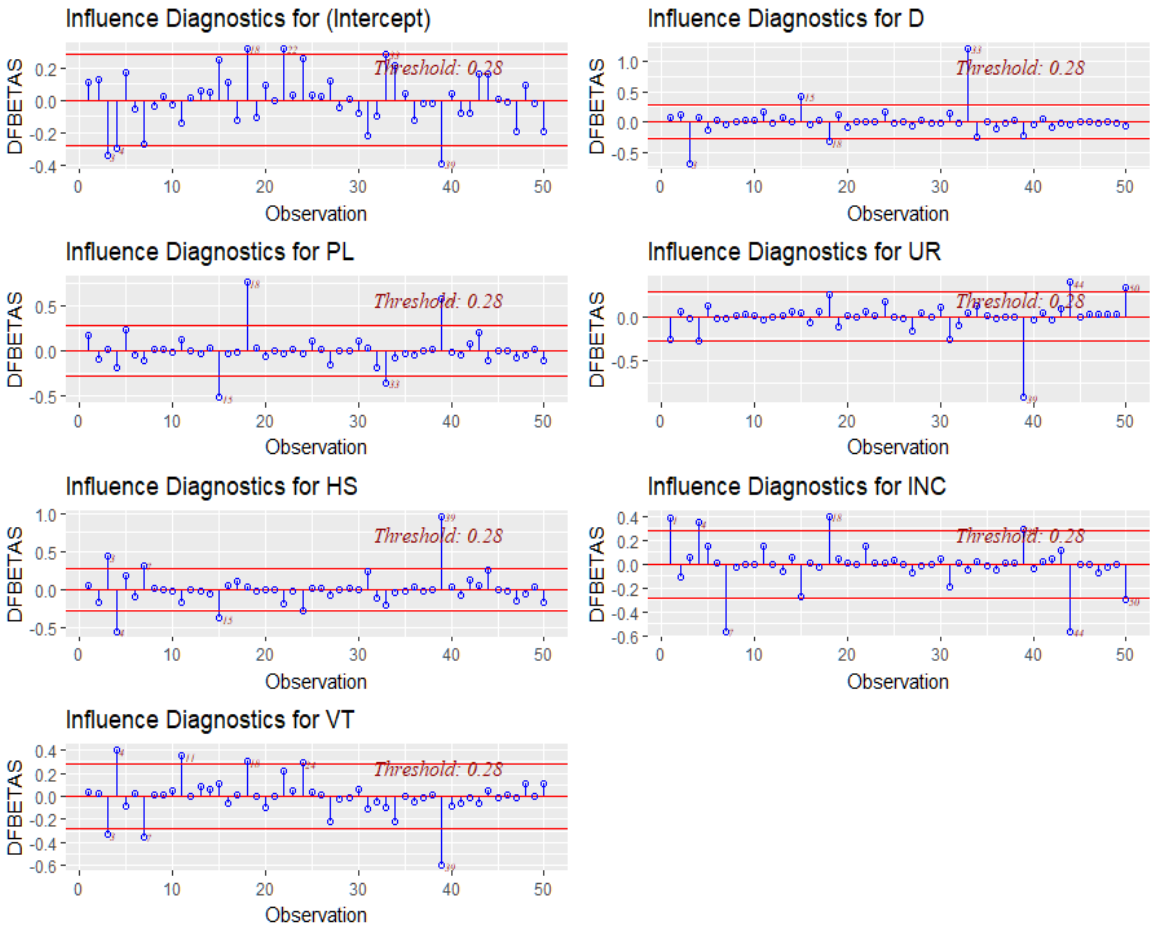
Table 5. Linear regression model summary for 1st problem



Graph 6. Residual diagnostic of the chose model for 1st problem



Graph 7. Influential point and outlier diagnostic of the chose model for 1st problem

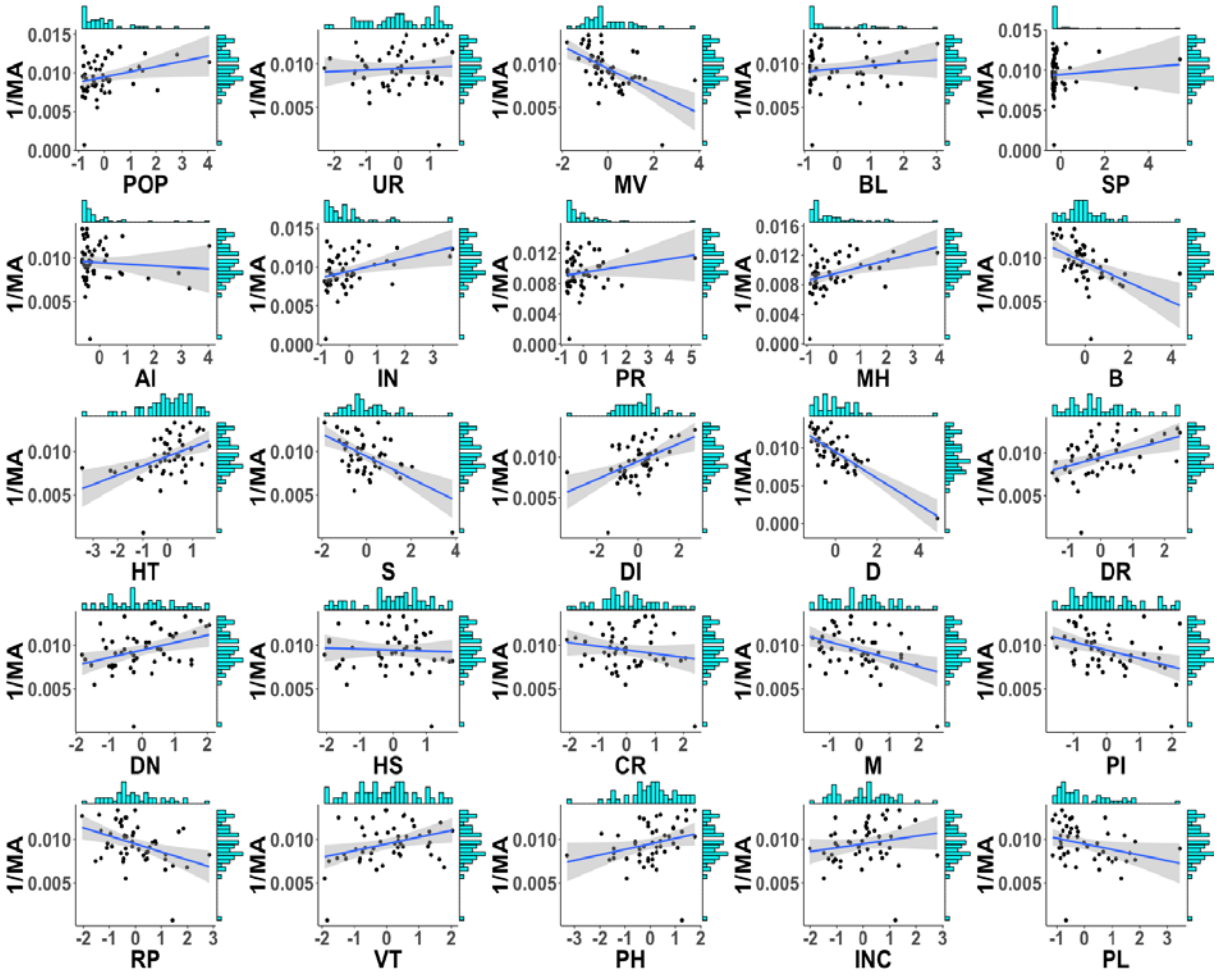


Graph 8. DFBETAs panel of the chose model for 1st problem

When doing the inference on the effect sizes predicted by our built model (Table 5), we discovered that the Per mil of population below poverty level (PL) and Per mil high school grads (HS) serve as the most and the least impactful independent variable respectively to Murder rate per 100,000 population (M), whilst other effect sizes were close the HS'. On the aspect of the sign of the effects, it made perfect sense that a higher Divorce rate per 10,000 (D), a higher PL, a higher Per mil of population living in urban areas (UR), a lower HS, and a lower % voting for presidential candidates among voting age population (VT) would stimulate a higher murder rate (M). However, the reasoning of a higher Per capita income expressed in 1972 dollars (INC), which also demonstrated a strong negative correlation to PL, would increase the M didn't come around naturally. Maybe because INC was like a mean instead of a median, which traditionally considered as a better representation for the mass' income. Either way, this issue should be brought to our respected client's full attention.

3.2. Find an appropriate linear regression model with MA as the dependent variable along with all other variables to choose from as independent variables.

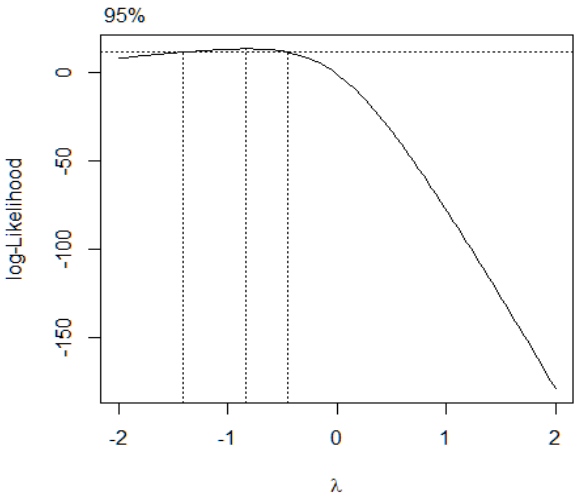
In the initial EDA, we checked the partial scatter plots of each candidate independent variable versus the 1/response like in Graph 9. Aside from several influential points for a few independent variables like D and PR, all other plots look decent enough, hinting the transformation of inverse the as-is response MA. We then carried out the model adequacy test by the Near-neighbor approach, and got a well below any common significance level p-value of 7.3e-4. It proved our initial guess that some transformation was indeed required on the as-is linear model.



Graph 9. Partial scatter plots in 2nd problem

To address this issue, we first tried to do a Boxcox power transformation on our response, the plot was shown in Graph 10. -1 was inside the 95% C.I. of the max log-likelihood function, hence we decided to do an inverse transformation on our response MA. The corresponding Near-neighbor approach model adequacy test rendered a p-value of 0.184, indicating merely inverting the MA would make the model adequate for the ensued procedure. This was also the very reason we started off by showing partial plots between the inverse of MA versus all other independent

variables. We then decided to do some additional transformation of ranking on Number of blacks (1000's) (BL), Number of Spanish speaking (1000's) (SP), and Number of Native Americans (1000's) (AI). And this procedure indeed improved the Near-neighbor approach model adequacy test with an updated p-value of 0.494.



Graph 10. Boxcox power transformation on the response plot in 2nd problem

Next, we proceeded with the variable selection from all candidates following a step-wise both-direction fashion evaluated by various different criteria. The results are listed below in Table 6. However the outputs was only D, B, INC and DI four variables with a moderate adjusted R^2 values and pretty abysmal C(p) values, we continued to seek a more fitting model. On the other hand, it should be noted that we have 25 independent variables so there were $2^{25} > 3.3E7$ possible models, which was too much for my gear to handle. After running several variables selections with different criteria and taking their sociological meanings into consideration, we decided to run a overall best subset model screening from 15 independent variables. However, as we could see, the best model remained unchanged, which was:

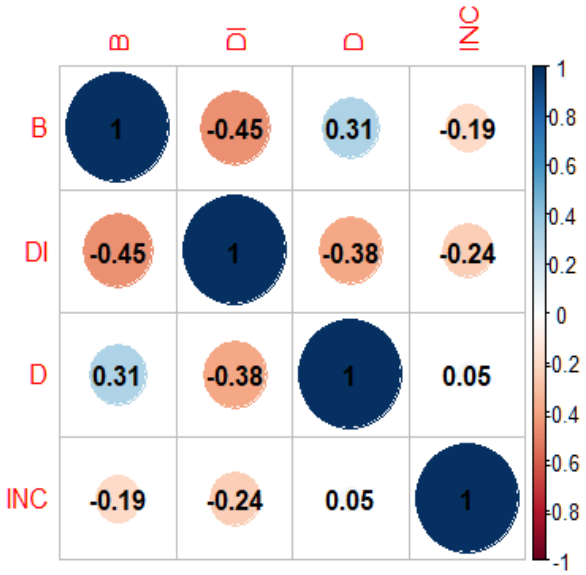
$$\frac{1}{MA} = D + B + INC + DI + \epsilon$$

And the correlation table was displayed in Graph 11.

```
## Stepwise Selection Method
##
## Candidate Terms:
##
## 1 . POP
## 2 . UR
## 3 . MV
## 4 . BL
```

```
## 5 . SP
## 6 . AI
## 7 . IN
## 8 . PR
## 9 . MH
## 10 . B
## 11 . HT
## 12 . S
## 13 . DI
## 14 . D
## 15 . DR
## 16 . DN
## 17 . HS
## 18 . CR
## 19 . M
## 20 . PI
## 21 . RP
## 22 . VT
## 23 . PH
## 24 . INC
## 25 . PL
##
## ----------------------------------------------------------------------------------------
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------
##                         Added/                   Adj.
## Step    Variable     Removed      R-Square    R-Square     C(p)         AIC         RMSE
## ----------------------------------------------------------------------------------------
##    1        D         addition       0.565       0.556    12.4790    -501.6771    0.0015
##    2        B         addition       0.639       0.623     4.6160    -508.9128    0.0014
##    3       INC        addition       0.669       0.648     2.5070    -511.3280    0.0014
##    4        DI        addition       0.702       0.675     0.1090    -514.5297    0.0013
```

Table 6. Step-wise variable selection by p-value for 2nd problem



Graph 11. Correlation matrix of the chose model for 2nd problem

In addition, we tested the chosen model's multicollinearity in Table 7. All variables' VIFs were well below 10 and condition indices were well below 30. This was not surprised at all given we only have four independent variables left. Thus, we could say for sure this chosen model didn't possess a prominent multicollinearity issue.

```
## Tolerance and Variance Inflation Factor
## ---------------------------------------
## # A tibble: 4 x 3
##   Variables Tolerance      VIF
##       <chr>     <dbl>    <dbl>
## 1         B 0.6808011 1.468858
## 2        DI 0.6334173 1.578738
## 3         D 0.8308413 1.203599
## 4       INC 0.8250008 1.212120
##
##
## Eigenvalue and Condition Index
## ------------------------------
##   Eigenvalue Condition Index intercept         B         DI            D
## 1  1.7699447        1.000000         0 0.1196873 0.13975880 0.1361593391
## 2  1.1539344        1.238481         0 0.1013843 0.02737701 0.0003319921
## 3  1.0000000        1.330393         1 0.0000000 0.00000000 0.0000000000
## 4  0.6883794        1.603489         0 0.1661382 0.10485948 0.8470174478
## 5  0.3877415        2.136529         0 0.6127901 0.72800472 0.0164912209
##           INC
## 1 0.003846077
## 2 0.556100422
## 3 0.000000000
## 4 0.019501905
## 5 0.420551596
```

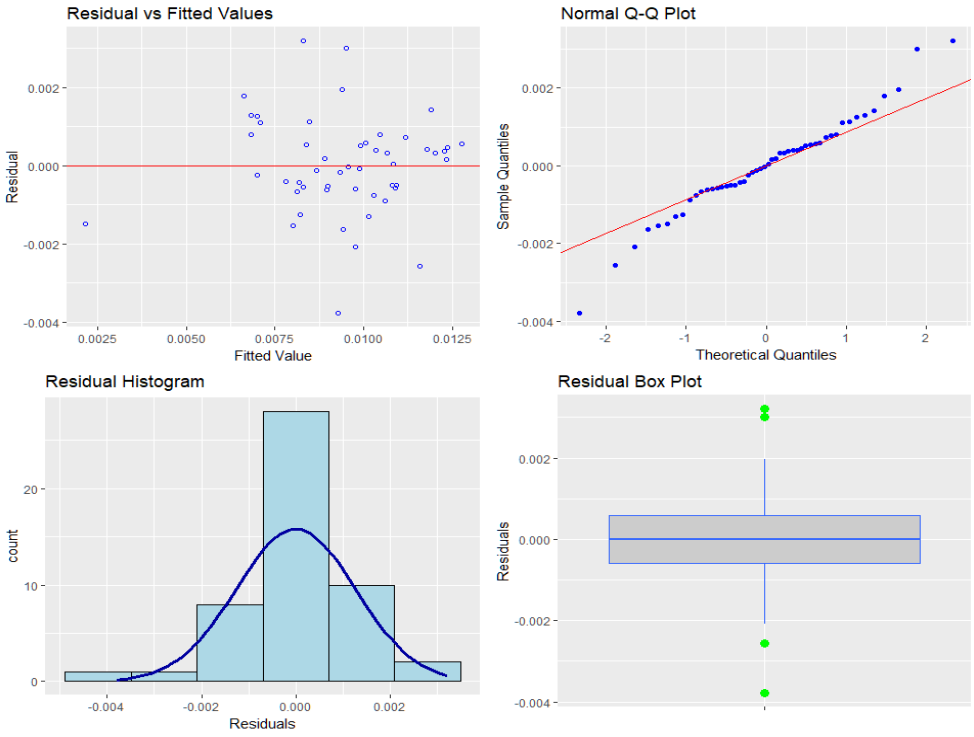
Table 8. Multicollinearity test for 2nd problem

Then, the chosen model's linear regression result was displayed in Table 8, including the model summary, ANOVA table and the parameter estimates including the 95% C.I.. It should be noted although we included all the interaction terms but none was significant, meaning all the independent variables tended to influence the response independently. The adjusted R^2 value was 0.675, demonstrating a fairly high level goodness-of-fit. From the residual diagnostic (Graph 12.), we could observe no clear pattern, indicating the validity of the linearity. The normality assumption was also supported by the qqplot and histogram, and further corroborated by S-W test (p-value:0.319) and K-S test (p-value:0.692). Meanwhile, form the influence measures (Graph 13.), we found observation 33 (NV) still stray furthest from the herd in both Cook's D and Leverage metric for the same reason stated in 3.1. Furthermore, DFBETAs (Graph 14.) measured the difference in each coefficient estimate with and without the influential point. From which, we could get senses on how much one observation has effected the estimate of a regression coefficient. Clearly, apart from the NV, observation 27 (NC) also showed huge influences on variable B, DI and INC. Observation 37 (OR) also showed huge

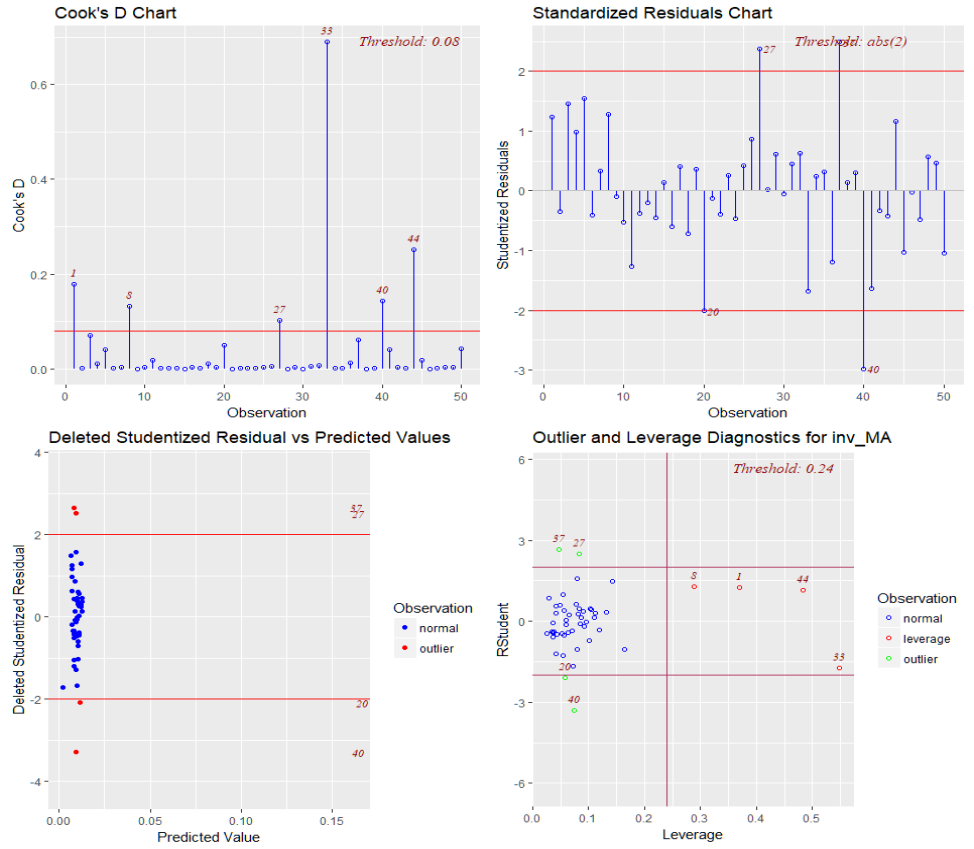
influences on variable DI. Observation 40 (SC) also showed huge influences on variable INC. Observation 44 (UT) also showed huge influences on variable B. On the other hand, we should expect 2.5 outliers on average out of our 50 observations. In practice, 4 were found (Graph 13.), and our dear clients should treat them with extra care. In particular, the "worst" one of them, possessing both the largest leverage and studentized residual was observation 40 (SC), with an alarming high studentized residual larger than 3. Nevertheless, one should provide more sociological reasoning if this outlier needed to be removed.

```
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.838       RMSE                 0.001
## R-Squared               0.702       Coef. Var           13.933
## Adj. R-Squared          0.675       MSE                  0.000
## Pred R-Squared          0.595       MAE                  0.001
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## -------------------------------------------------------------------
## Regression      0.000         4          0.000     26.48    0.0000
## Residual        0.000        45          0.000
## Total           0.000        49
## -------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------
## (Intercept)     0.009         0.000                 50.751    0.000     0.009     0.010
##           B     0.000         0.000       -0.142    -1.442    0.156    -0.001     0.000
##          DI     0.001         0.000        0.227     2.221    0.031     0.000     0.001
##           D    -0.001         0.000       -0.634    -7.097    0.000    -0.002    -0.001
##         INC     0.001         0.000        0.251     2.805    0.007     0.000     0.001
## ----------------------------------------------------------------------------------------------
```

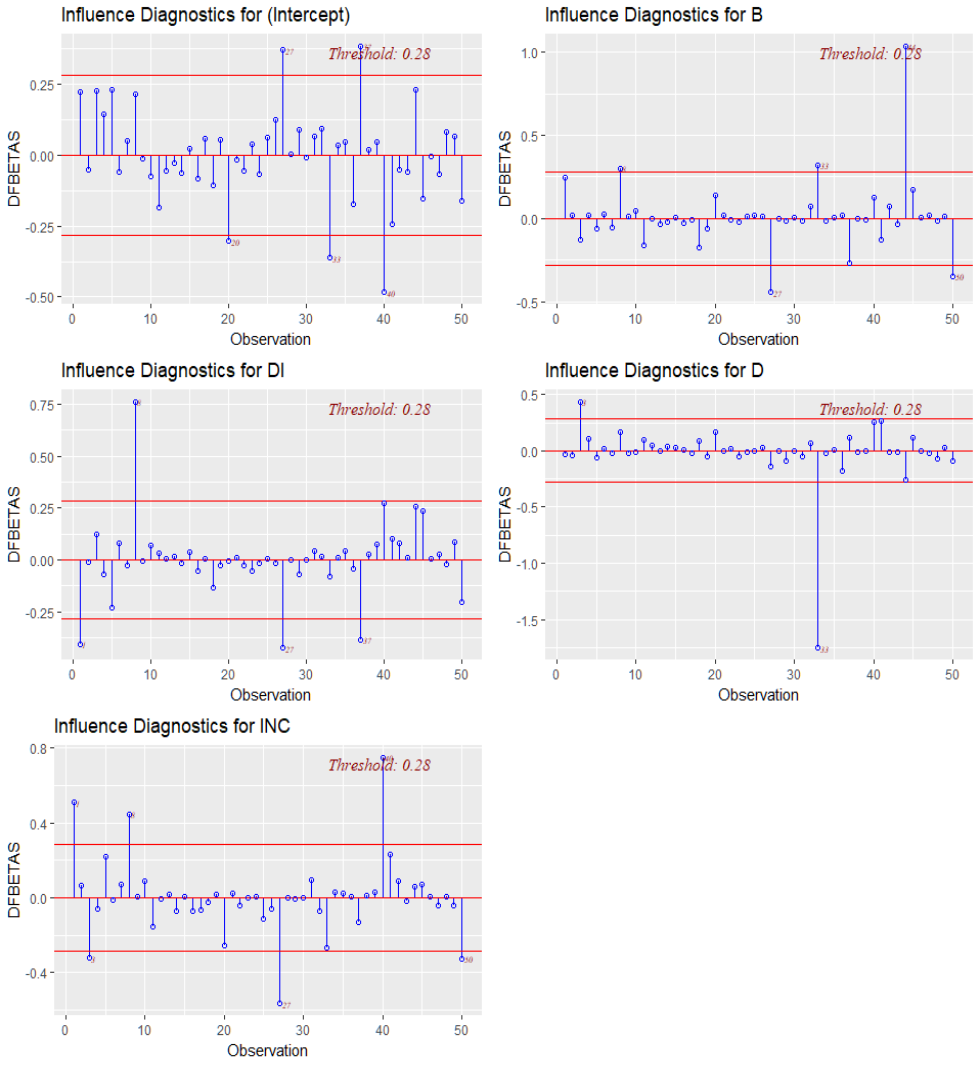
Table 9. Linear regression model summary for 2nd problem



Graph 12. Residual diagnostic of the chose model for 2nd problem



Graph 13. Influential point and outlier diagnostic of the chose model for 2nd problem



Graph 14. DFBETAs panel of the chose model for 2nd problem

When doing the inference on the effect sizes predicted by our built model (Table 8), we discovered that the Divorce rate per 10,000 (D) serve as the most impactful independent variable to the inverse of Marriage rate per 10,000 (1/MA), whilst all of the rest effect sizes were close to each other, which seemed natural. On the aspect of the sign of the effects, it made perfect sense that a higher Births per 1000 (B), would stimulate a higher Marriage rate per 10,000 (MA), though it was not significant. It also made some sense that a lower Per capita income expressed in 1972 dollars (INC) would stimulate a higher Marriage rate per 10,000 (MA), since folks richer classes probably have more choices in their life styles rather than getting married. Moreover, we found that a lower Death rate from diabetes, 1978, per 100,000 (DI) would stimulate a higher Marriage rate per 10,000 (MA), maybe our respected client could infer the cause behind this phenomenon. At last, it was a bit shocking to discover that a higher Divorce rate per 10,000 (D) would also stimulate a higher Marriage rate per 10,000 (MA). Perhaps there existed some physiological explanations on a certain group of people tend to rush to all marriage related

decisions, regardless of marrying someone of divorcing someone. Either way, this issue should be brought to our respected client's full attention.

4. Conclusions

For this project, in order to construct both accurate and interpretable linear regression models on the desired responses, we first conducted appropriate transformations on the response variable and predictors until it could pass the Near-neighbor approach based model adequacy test. Next, we performed step-wise both-direction variable selections with p-values, AIC, adj-R^2, and C(p) as criteria for a comprehensive consideration. Then, the selected independent variables needed to pass the multicollinearity test. Eventually, linear regression models were constructed and ensued by residual and influence diagnostics. As a result, we discovered that higher D, PL, UR, INC plus lower HS and VT will likely lead to increases in the murder rate (M) in our first model with an adj-R^2 of 0.843. Meanwhile, the second model revealed that higher D plus lower B, DI and INC will likely lead to increases in the marriage rate (MA) with an adj-R^2 of 0.675. Moreover, we offered our amateur two-cents towards the interpretation of the models. In the last but not the least, we highly encouraged special attention should be allocated to the outliers and influential points we found by our respected clients.