Homework 4

Muhe Xie

NetID: mx419

Exploration of CO2 Emissions Data of China

Data Overview

Global warning is viewed as a great threat to human's food supply and living conditions and CO2 is regarded as one of the major gases that are responsible for global warming. Thus people try to the monitor and restrict the emissions of CO2 for countries all over the world. As a developing country which is currently developing rapidly, China would be an interesting target to study.

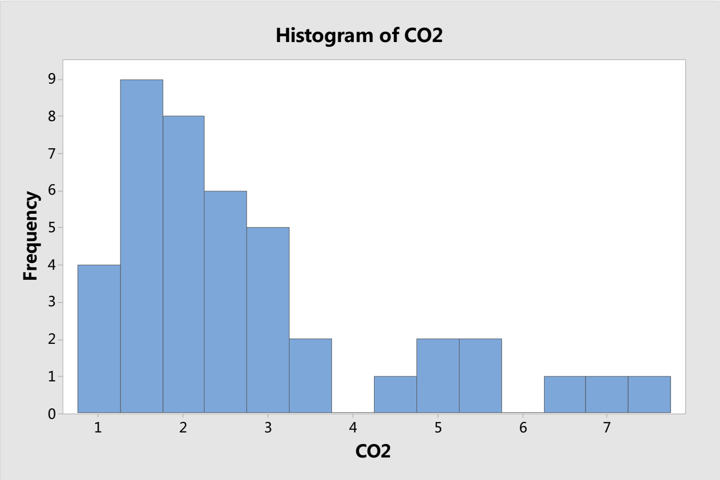
In this report, CO2 emissions (metric tons per capita, use CO2 for short in the following context) is the response variable. And initial three predicting variables are listed below:

1 Urban population (% of total, use Urban_Pop for short)
2 Arable land (% of land area)
3 Electric power consumption (kWh per capita, use Ele_Csp for short)

Since these three variables are related with urbanization, land cover and energy consumption, it would be reasonable to guess that they can be appropriate predictors for CO2 emissions. And also, I create a variable Time which is the order of data, and another variable Time Square which is derived from Time.

The data is collected from 1971 to 2012 (42 time points, which is the longest consecutive sub-sequence in the available data) and the detailed definitions of theses variables are attached at the end of this report. The data is downloaded from World Bank.

First, look at the response data below:

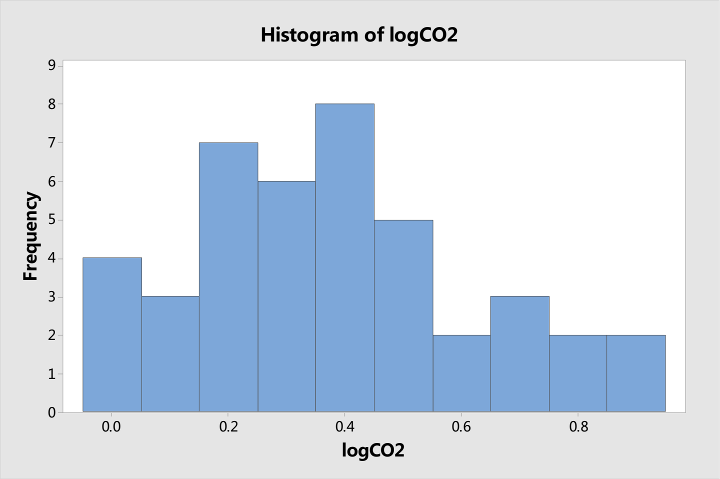


The basic statistics is shown below:

Stats

```
Variable    N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
CO2        42   0  2.806    0.261  1.693    1.042  1.561   2.280  3.192    7.419
```

The histogram does show a pattern of 'long right tail', thus I decide to take log (10 base) to the response data, check the new histogram below:

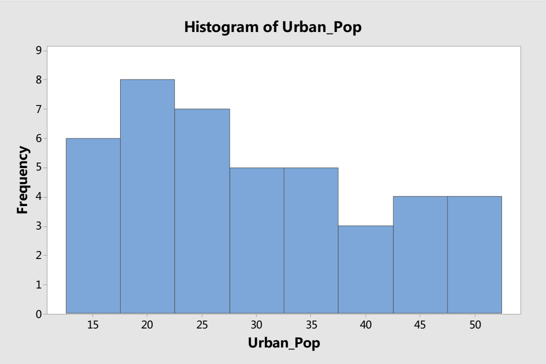


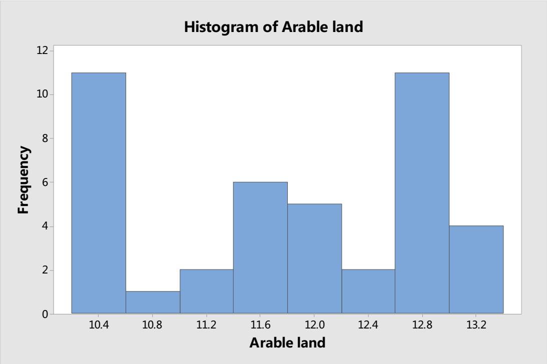
Then look at the predicting variables.

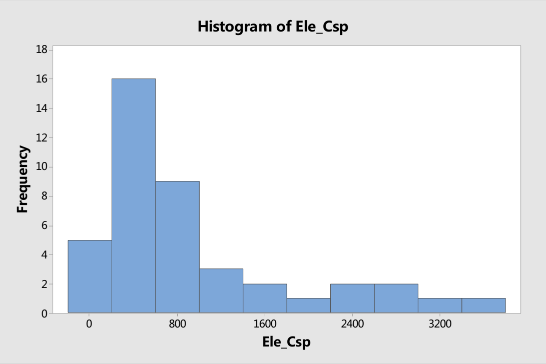
Basic Stats:

```
Variable      N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
Urban_Pop    42   0   29.97     1.69  10.92    17.18   19.93   27.76   38.76    51.89
Arable land  42   0  11.780    0.160  1.040   10.244  10.519  11.893  12.773   13.301
Ele_Csp      42   0     951      142    919      152     285     577    1242     3475
```

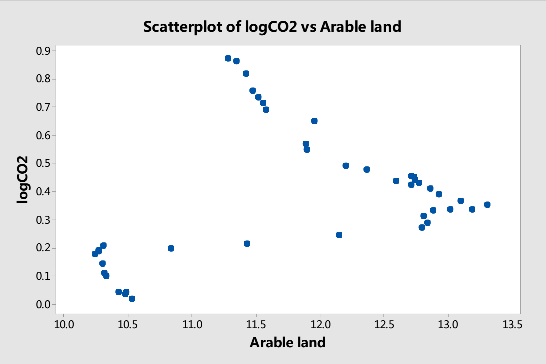
The histograms are shown below:

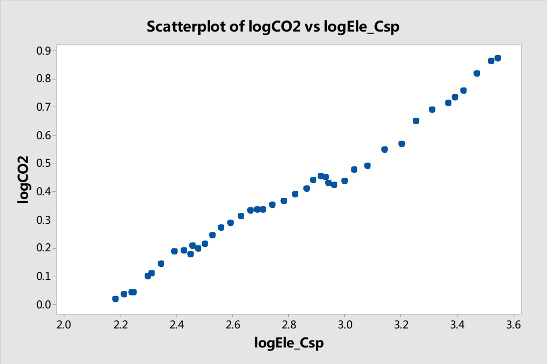


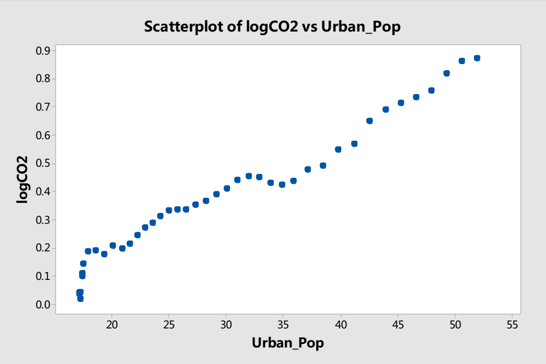




Again the variable Ele_Csp has a long right tail, I will take log for that too.
Then we take a look at the scatter plots:



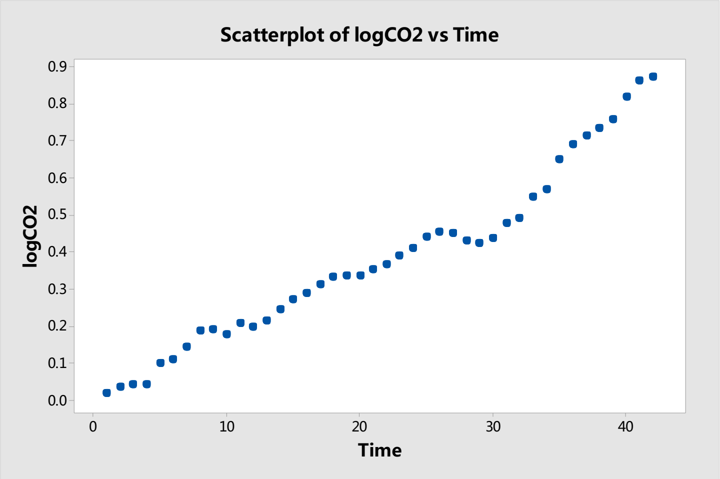




Apparently, logEle_Csp and Urban_Pop have strong linear relationships with the response logCO2 while the Arable land seems to be not much relevant and has non-constant variance (we will see it later). And the scatter plot of Arable land seems to have the issue of non-constant variance.

Regression Analysis

Before we start the regression analysis. It won’t hurt to see the response vs time plot, which is shown below:



Apparently, the plot of logCO2 vs Time shows that the logCO2 increases steadily with Time, thus I would also include Time and Time Square in our model at the very beginning (detrending), which might be helpful to address autocorrelation if there is any.

The initial regression on all variables, Time and Time Square is shown below:

Regression Analysis: logCO2 versus Arable land, Urban_Pop, logEle_Csp, Time Square, Time

```
Analysis of Variance


Source          DF   Adj SS    Adj MS  F-Value  P-Value
Regression       5  2.25210  0.450420  1435.40    0.000
  Arable land    1  0.00285  0.002849     9.08    0.005
  Urban_Pop      1  0.00248  0.002481     7.91    0.008
  logEle_Csp     1  0.05134  0.051344   163.62    0.000
  Time Square    1  0.00126  0.001255     4.00    0.053
  Time           1  0.01389  0.013887    44.25    0.000
Error           36  0.01130  0.000314
Total           41  2.26340



Model Summary


        S    R-sq  R-sq(adj)  R-sq(pred)
0.0177143  99.50%     99.43%      99.29%



Coefficients
```

```
Term              Coef   SE Coef  T-Value  P-Value      VIF
Constant        -2.966     0.283   -10.47    0.000
Arable land    0.02059   0.00683     3.01    0.005     6.60
Urban_Pop      -0.0314    0.0112    -2.81    0.008  1947.92
logEle_Csp       1.524     0.119    12.79    0.000   298.64
Time Square   0.000402  0.000201     2.00    0.053  1560.41
Time          -0.02141   0.00322    -6.65    0.000   203.68


Regression Equation

logCO2 = -2.966 + 0.02059 Arable land - 0.0314 Urban_Pop + 1.524 logEle_Csp
         + 0.000402 Time Square - 0.02141 Time
```

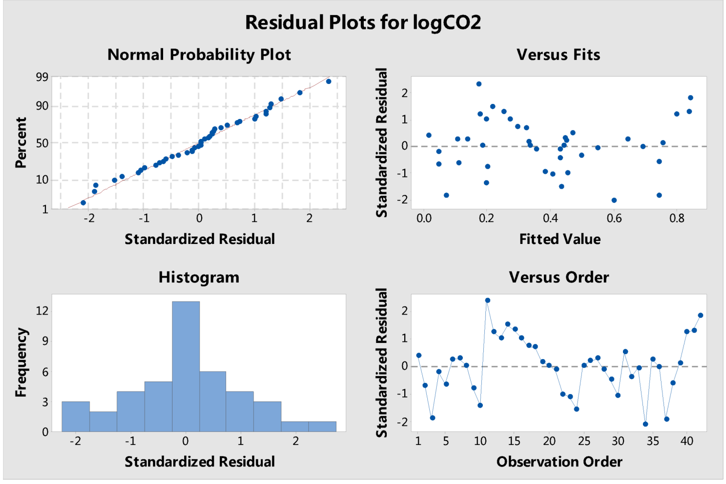
The overall regression is very statistically significant with the large F-value and small P value (close to 0). The R Square also looks impressively good (99.50%). It means more than 99% of the response is associated with the predictors. All coefficients of predictors are statistically significant except Time Square is a little marginal. And it seems that multicollinearity has become a problem here with large VIF. We should do best subsets later. The standard error of the estimate is about 0.018, indicating a roughly 95% prediction interval for the logCO2 is ± 0.036 of our best guess, which means the prediction interval for CO2 emissions is as high as 1.09 ($10^{0.036}$) of our best guess and as low as 92% ($10^{-0.036}$) of our best guess.

According to the Regression Equation, although the coefficients of Urban_Pop and Time are expected to be positive, they are actually negative. That would not be surprising since we have multicollinearity issue in our model ($1/(1 - R^2_{\text{model}}) \approx 200$, and we have several VIF larger than 200), which makes the model not very stable and robust. I prefer not interpreting the coefficients here because of the high multicollinearity here. I will do interpretation in the following simpler models. We can see the correlation form below to see the multicollinearity:

Correlation: logCO2, Urban_Pop, logEle_Csp, Arable land, Time

```
                  logCO2    Urban_Pop   logEle_Csp  Arable land
Urban_Pop          0.983
logEle_Csp         0.993        0.993
Arable land        0.337        0.303        0.366
Time               0.978        0.980        0.994        0.445
```

The residual plots are shown below:



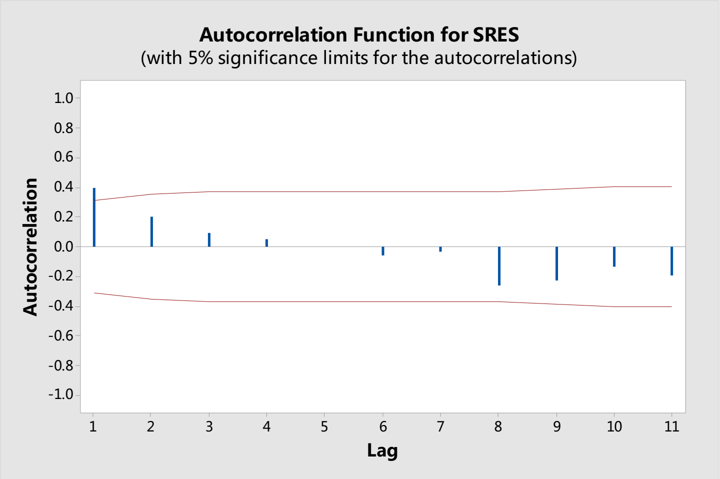
According to the residual plot, the normality assumption looks almost fine. There is structure and non-constant variance in the residual vs fitted value plot. Also, there seems to be some potential outliers and leverage points (which we will address later) in the plots. And we can see there might be autocorrelation problem according to the order plot. The D-W test is shown below:

```
Durbin-Watson Statistic


Durbin-Watson Statistic =  1.11255
```

After checking the D-W statistic form, we can roughly reject the null hypothesis at 0.05 level. So there is probably an autocorrelation issue. Moreover, we can see the run test and ACF plot:

Runs Test: SRES

```
Runs above and below K = 0


The observed number of runs = 13
The expected number of runs = 21.8095
23 observations above K, 19 below
P-value = 0.005
```



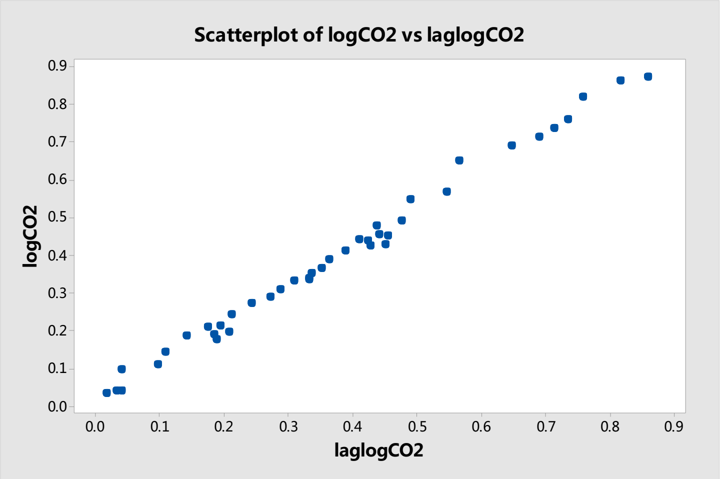
The runs test and ACF plot shows that there is autocorrelation issue here (lag 1 is 5% significant). Thus we are not done yet with auto-correlation.

In conclusion, there are three major problems we are facing. First, D-W test, Runs test and ACF plot shows that there might be an issue of autocorrelation. The possible solution is adding a lagged response in the predicting variables. Second, the potential outliers and leverage points. The possible solution is doing regression diagnostics and looking at these unusual points carefully to find out how to deal with them. The third problem is multicollinearity. The solution is using best subsets method to see if we can eliminate some redundant variables. So I need to decide what should we do first. Here I choose to deal with autocorrelation issue first. The reasons are listed below:

1 After addressing the autocorrelation issue, the result of best subsets method would be more reliable according to the lecture. Thus I prefer to address autocorrelation before I doing model selection.

2 After adding a lagged variable (and we may not need variable Time), the model is actually different and the previous outliers or leverage points may change. Since I only have 42 data points, I prefer to be very cautious to throw out the data points. Thus I prefer to address the autocorrelation first and do regression diagnostics later.

The possible solution is adding a lagged response variable as a predictor. The plot of logCO2 vs laglogCO2 is shown below:



It looks quite reasonable to include the lag variable in the predictors. And since we have the lagged response variable, the variable Time and Time Square would probably not be needed in the model given that the variable Time is highly correlated with other variables in the model (0.98 with Urban_Pop, 0.994 with logEle_Csp) and it does not address the autocorrelation problem. So it is a reasonable choice to try to do the regression based on laglogCO2 and other predictors but without Time and Time Square (will consider adding it back if it does not work) to see the result first:

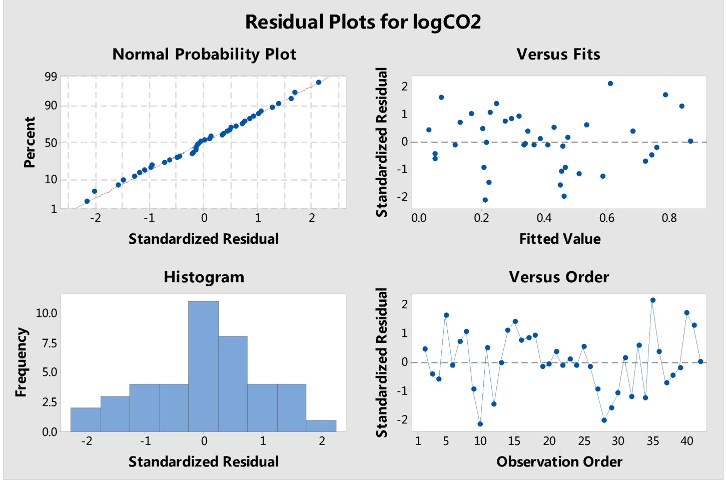
Regression Analysis: logCO2 versus laglogCO2, logEle_Csp, Urban_Pop, Arable land

```
Analysis of Variance

Source          DF    Adj SS     Adj MS  F-Value  P-Value
Regression       4   2.11513   0.528781  1570.86     0.000
  laglogCO2      1   0.01492   0.014916    44.31     0.000
  logEle_Csp     1   0.00228   0.002280     6.77     0.013
  Urban_Pop      1   0.00059   0.000590     1.75     0.194
  Arable land    1   0.00164   0.001641     4.88     0.034
Error           36   0.01212   0.000337
Total           40   2.12724

Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0.0183472  99.43%     99.37%      99.26%


Coefficients
```

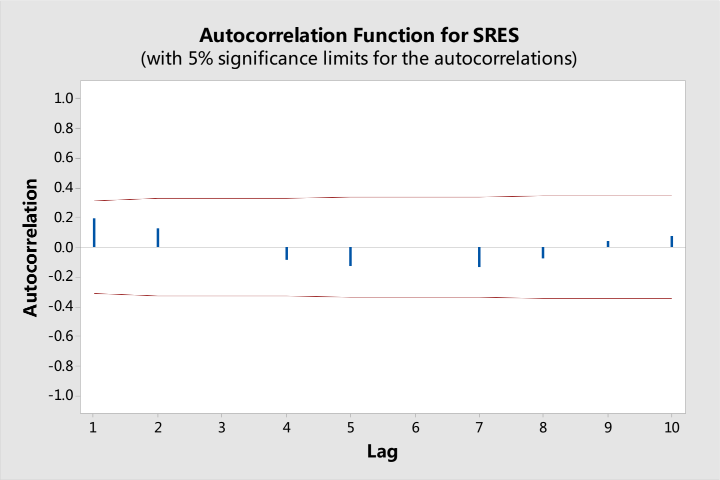
```
Term              Coef  SE Coef  T-Value  P-Value     VIF
Constant        -0.528    0.208    -2.53    0.016
laglogCO2        0.685    0.103     6.66    0.000   63.51
logEle_Csp       0.316    0.122     2.60    0.013  272.64
Urban_Pop     -0.00422  0.00318    -1.32    0.194  142.29
Arable land   -0.00817  0.00370    -2.21    0.034    1.74


Regression Equation

logCO2 = -0.528 + 0.685 laglogCO2 + 0.316 logEle_Csp - 0.00422 Urban_Pop
         - 0.00817 Arable land
```

The overall regression is very statistically significant with the large F-value and small P value (close to 0). The R Square also looks impressively good (99.4%). The standard error of the estimate is 0.018, indicating a roughly 95% prediction interval for the logCO2 is ± 0.036 of our best guess, which means the prediction interval for CO2 emissions is as high as 1.09 ($10^{0.036}$) of our best guess and as low as 92% ($10^{-0.036}$) of our best guess.

Because of the multicollinearity, it would not be reliable to interpret the coefficients here (because the condition "given all other variable are held fixed" would be less meaningful when there is multicollinearity). Here we can see the problem of multicollinearity more clearly. In fact, according to the scatter plot, the Urban_Pop is strongly positively linearly related with logCO2. But the coefficient here is negative which is not reasonable. So simply interpreting the model by the terminology "holding other predictors fixed" would not be meaningful. Which will drive me to do model selection in the next step.

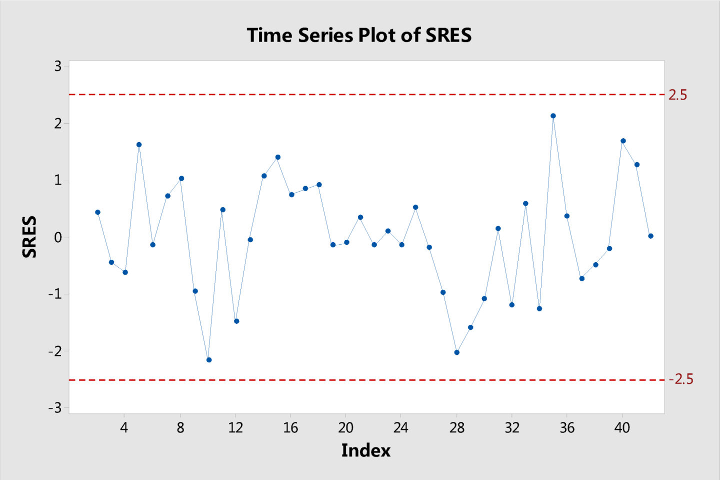


According to the residual plot, the residual vs fits plot shows less structure and the normality looks roughly OK except that there might be some leverage points. And the auto-correlation issue seems to be addressed. Look at the autocorrelation condition. The ACF plot and run test are shown below:



```
Runs test for SRES_1


Runs above and below K = 0


The observed number of runs = 23
The expected number of runs = 21.4878
20 observations above K, 21 below
P-value = 0.632
```

It seems that we have successfully addressed the autocorrelation problem. Then there are two options to do, check the diagnostics first or do model selection first. If we just take a look at the standardized residual below, it looks roughly fine. And similarly, I prefer to do model selection first and do more detailed diagnostics about unusual points later.



Best Subsets Regression: logCO2 versus Arable land, Urban_Pop, ...

```
Response is logCO2
41 cases used, 1 cases contain missing values

                                              A
                                              r   l
                                              a U o l
                                              b r g a
                                              l b E g
                                              e a l l
                                                n e o
                                              l _ _ g
                                              a P C C
             R-Sq    R-Sq  Mallows            n o s O
Vars  R-Sq  (adj)  (pred)       Cp          S d p p 2
   1  99.2   99.2    99.1     12.4  0.020645        X
   1  98.4   98.4    98.2     62.2  0.029259      X
   2  99.4   99.3    99.2      5.9  0.019029      X X
   2  99.3   99.3    99.2      8.6  0.019643    X   X
   3  99.4   99.4    99.3      4.8  0.018533  X   X X
   3  99.4   99.3    99.2      7.9  0.019284    X X X
   4  99.4   99.4    99.3      5.0  0.018347  X X X X
```

The best subsets method shows that the R-square is as high as 99.2 even with only one predictor (laglogCO2). So for the prediction point of view, actually one predictor model would be a possible solution. But since I am interested in what factor will affect the CO2 emissions. I prefer include at least one of the initial variables in the model rather than just include the lag variable. Thus there are two choices: The 2-predictor model (logEle_Csp and laglogCO2) and the 3-predictor model (logEle_Csp, laglogCO2 and Arable land). The 2-predictor one maximizes R-sq with least number of predictors and the Cp value is relatively low. The 3-predictor one minimizes Cp. Here I prefer the 2-predictor models and there are three reasons:

(1) The Arable land is close to marginally significant. And if I do a regression based on this 3-predictor model, the coefficients of the result are like below, indicating that the Arable land is not statistically significant.

```
Coefficients
Term              Coef  SE Coef  T-Value  P-Value    VIF
Constant        -0.304    0.123    -2.47    0.018
laglogCO2       0.7289   0.0984     7.41    0.000  56.90
logEle_Csp      0.1733   0.0563     3.08    0.004  57.41
Arable land   -0.00528  0.00302    -1.75    0.089   1.13
```

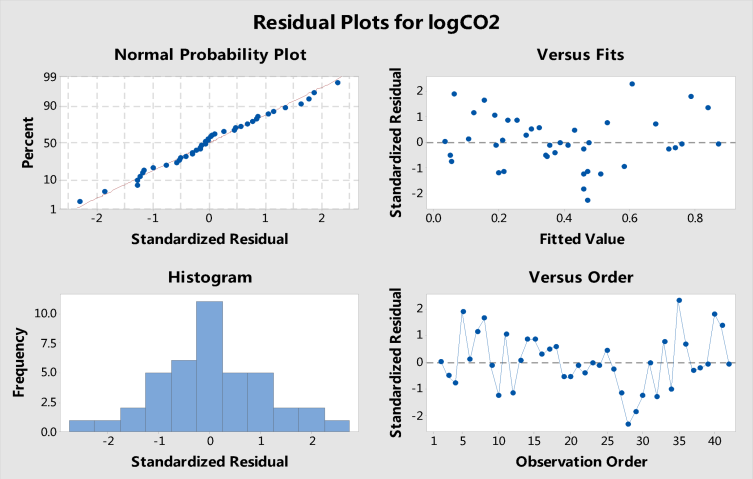
(2) I will prefer simpler model when the prediction power is close.
(3) Although the 3-predictor model minimizes Cp, by computing the MCp Value below, the 2-predictor model has actually smaller MCp.

Num of Predictors	MCp
2	1.59
3	1.74

Thus, we do a regression analysis on the 2-predictor model. The result is shown below:

Regression Analysis: logCO2 versus laglogCO2, logEle_Csp

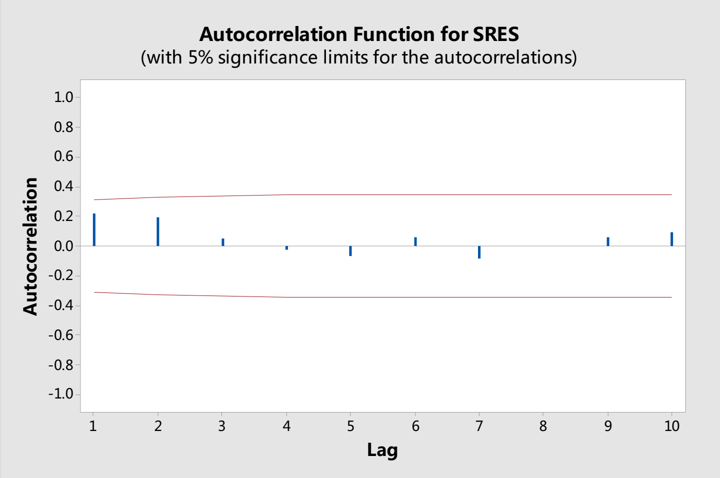
```
Analysis of Variance


Source        DF   Adj SS   Adj MS  F-Value  P-Value
Regression     2  2.11348  1.05674  2918.39    0.000
  laglogCO2    1  0.01963  0.01963    54.20    0.000
  logEle_Csp   1  0.00286  0.00286     7.91    0.008
Error         38  0.01376  0.00036
Total         40  2.12724



Model Summary


        S    R-sq  R-sq(adj)  R-sq(pred)
0.0190289  99.35%     99.32%      99.23%


Coefficients


Term          Coef  SE Coef  T-Value  P-Value    VIF
Constant    -0.338    0.125    -2.71    0.010
laglogCO2    0.742    0.101     7.36    0.000  56.59
logEle_Csp  0.1615   0.0574     2.81    0.008  56.59


Regression Equation


logCO2 = -0.338 + 0.742 laglogCO2 + 0.1615 logEle_Csp
```



In this model, both variables are statistically significant and the model is statistically significant. The R Square still looks impressively good (99.35%). The standard error of the estimate is 0.019, indicating that the 95% prediction interval for CO2 emissions is as high as 1.09 ($10^{0.038}$) of our best guess and as low as 92% ($10^{-0.038}$) of our best guess. The residual plot reveals some structure and non-constant variance. The normality assumption seems to be satisfied.

By computing $1/(1 - R^2_{\text{model}}) = 154$, the VIFs of the coefficients look fine. But given the multicollinearity between laglogCO2 and logEle_Csp, interpreting the coefficients marginally will not be very reliable, but I will just do that here anyway. The coefficient of laglogCO2 says that given electric power consumption per capita, one percent higher CO2 emission last year is associated with a 0.74% increase in CO2 emission this year. The coefficient of logEle_Csp says that given the last year's CO2 emission held fixed, one percent increase in electric power consumption per capita is associated with an 0.16% increase in CO2 emission this year. The constant has no direct interpretation. And we should not trust this interpretation too much.

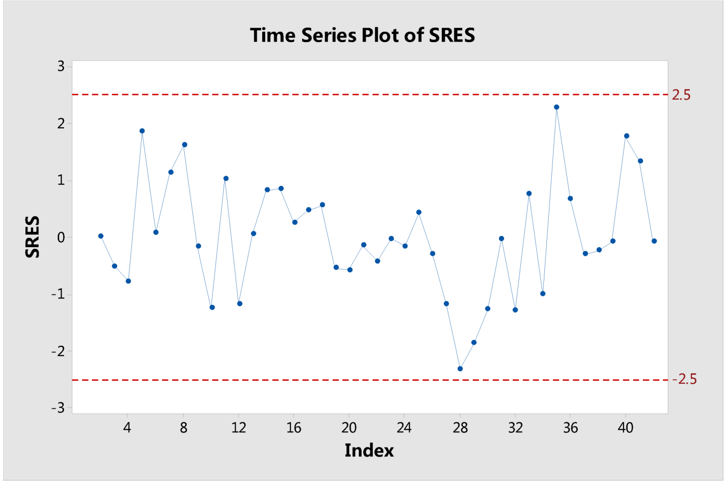
Then look at the autocorrelation condition. The ACF plot and the run test are shown below:

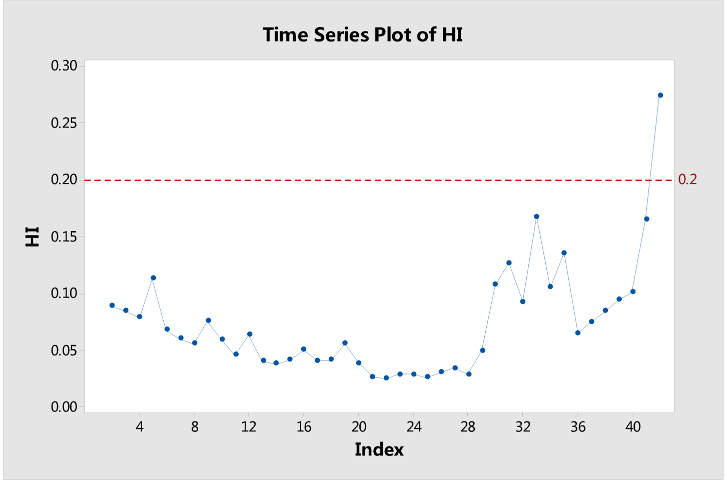


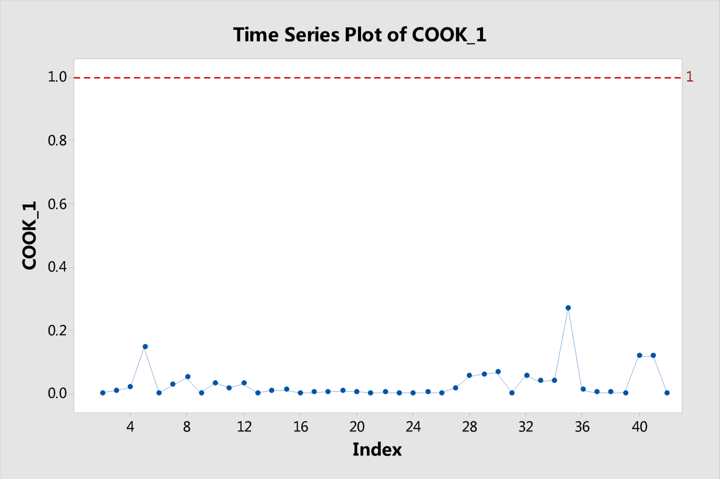
Runs Test: SRES_1

```
Runs above and below K = 0
The observed number of runs = 18
The expected number of runs = 20.9024
17 observations above K, 24 below
P-value = 0.344
```

The ACF plot and runs test says that there is no obvious autocorrelation issue in this model. The diagnostics are shown below:







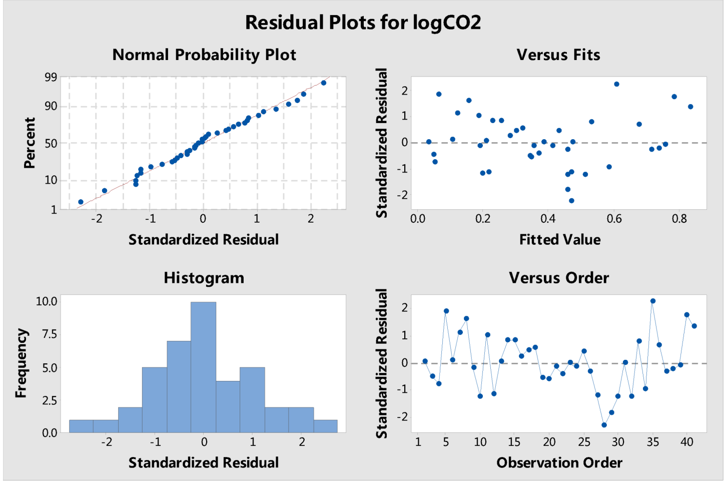
The SRES plot looks roughly fine but the Hii plot indicates an obvious leverage point which is Point 42 which refers to year 2012 which is the last year of my data series. It actually makes sense because the lagged response (or just the response) and the variable logEle_Csp increase with time. Thus it is possible that the beginning of the series or the end of the series will more likely to be the leverage point. We can see that point 42 has the largest laglogCO2 and logEleCsp. Thus this point in our model will be a potential problem since it will greatly affect the regression line. And I create a new indicator variable Year2012 in the model to omit the 42th observation and then rerun the regression.

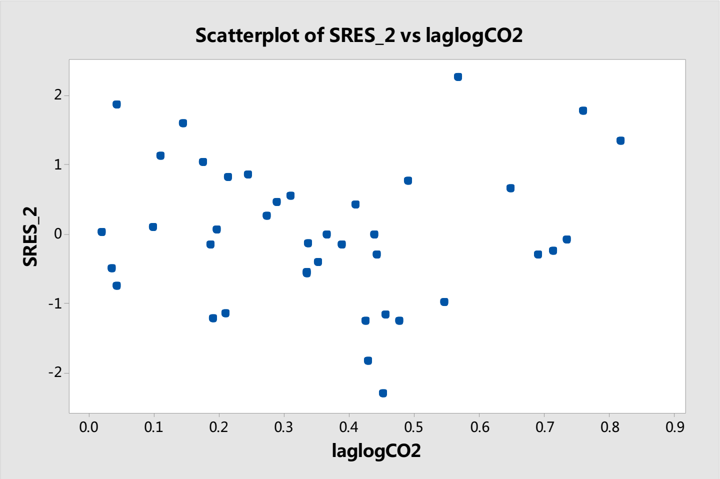
```
Analysis of Variance
Source          DF    Adj SS    Adj MS   F-Value  P-Value
Regression       3   2.11349  0.704496   1894.74    0.000
  laglogCO2      1   0.01618  0.016177     43.51    0.000
  logEle_Csp     1   0.00236  0.002363      6.35    0.016
  Year2012       1   0.00000  0.000003      0.01    0.935
Error           37   0.01376  0.000372
Total           40   2.12724

Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0.0192825  99.35%     99.30%           *


Coefficients

Term           Coef  SE Coef  T-Value  P-Value    VIF
Constant     -0.333    0.137    -2.44    0.020
laglogCO2     0.746    0.113     6.60    0.000  69.40
logEle_Csp   0.1595   0.0633     2.52    0.016  66.85
```

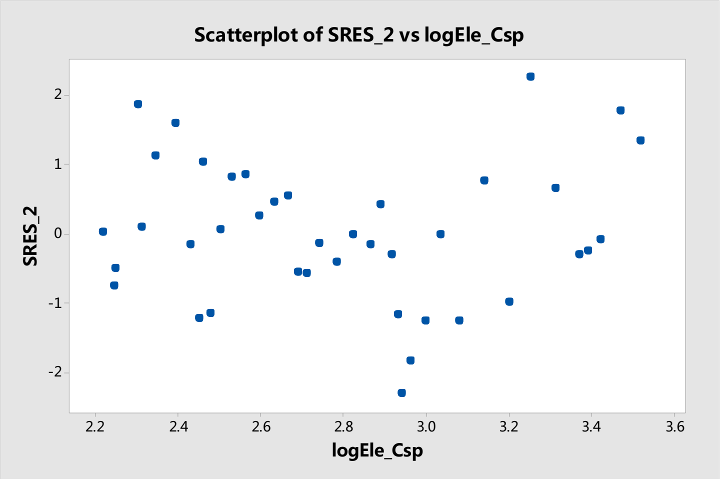
```
Year2012     -0.0019    0.0226     -0.08     0.935    1.34


Regression Equation

logCO2 = -0.333 + 0.746 laglogCO2 + 0.1595 logEle_Csp - 0.0019 Year2012
```

In this model, both variables are still statistically significant and the model is highly statistically significant. The R Square still looks impressively good (99.35%) and does not change much. The standard error of the estimate is 0.019, indicating that the prediction interval for CO2 emissions is as high as 1.09 (10 $^{0.038}$) of our best guess and as low as 92% (10 $^{-0.038}$) of our best guess.

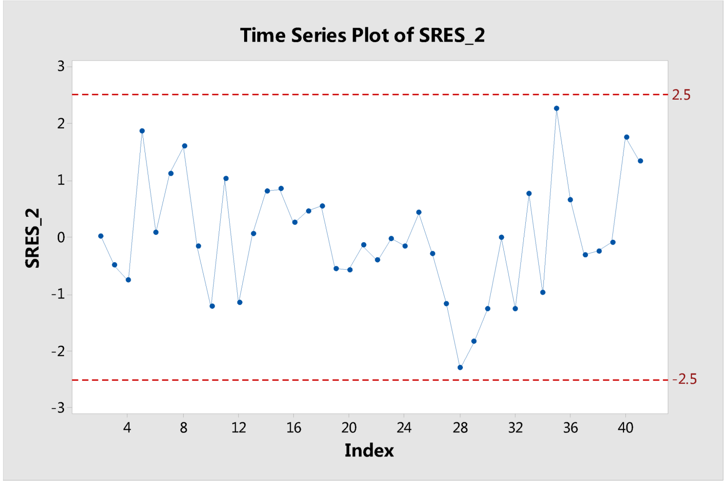
By computing $1/(1 - R^2_{model}) = 154$, the VIFs of the coefficients look fine. The coefficients do not change much. Again I will interpret the coefficients although we should not put too much weight on it because of multicollinearity. The coefficient of laglogCO2 says that given electric power consumption per capita, one percent more CO2 emission last year is associated with a 0.75% increase in CO2 emission this year. The coefficient of logEle_Csp says that given the last year's CO2 emission held fixed, one percent increase in electric power consumption per capita is associated with an 0.16% increase in CO2 emission this year. The constant has no direct interpretation.

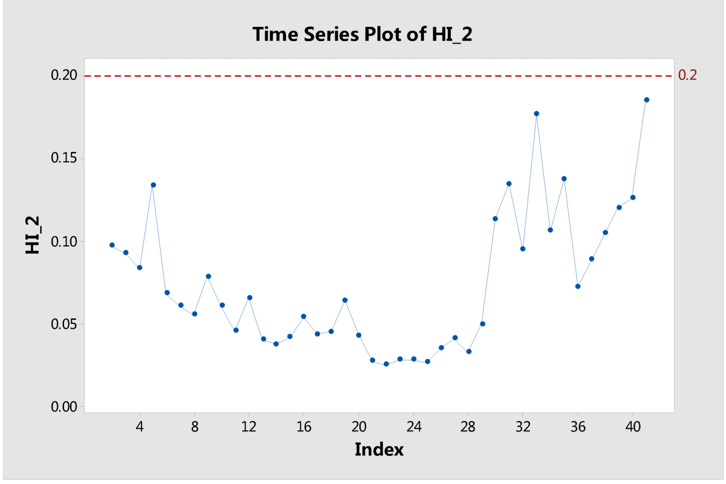




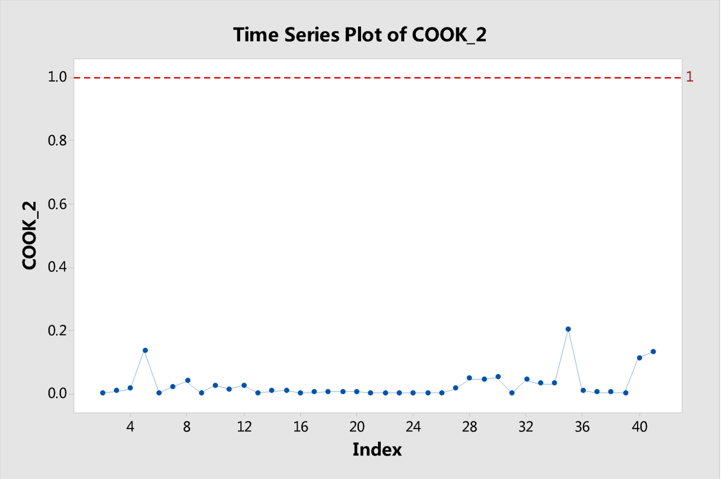


The residual plot looks good overall except that there is still some structure and a little non-constant variance in the residual vs fitted value plot. And by checking the plots of residual vs predictor, this problem has been more obvious. The normality assumption seems to be satisfied. And then I check the diagnostics again:

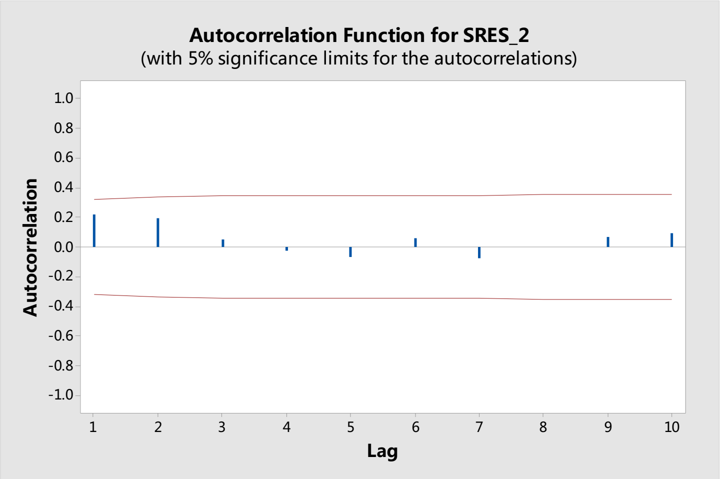




(omit the 42th point since it is coded with Hii equals 1)



The sres plot shows that all the standardized residuals are within [-2.5,2.5], and the Hii values are below the flag line. Cook's distance plot also seems to be fine.
And double check the autocorrelation below:



```
Runs test for SRES_4


Runs above and below K = 0


The observed number of runs = 17
The expected number of runs = 20.2
16 observations above K, 24 below
P-value = 0.285
```

The run test and ACF plot shows that the autocorrelation problem does not appear again. After addressing the leverage point issue, we need to re-run the best subset model, the result is shown below (here we already force the indicator variable in the model):

```
Response is logCO2
41 cases used, 1 cases contain missing values
```

```
                                                      A
                                                l     r
                                                o l U a
                                                g a r b
                                                E g b l
                                                l l a e
                                                e o n
                                                _ g _ l
                                                C C P a
Total         R-Sq    R-Sq  Mallows             s O o n
 Vars  R-Sq  (adj)  (pred)       Cp          S  p 2 p d
    2  99.2   99.2       *     11.6  0.020596     X
    2  98.6   98.5       *     51.5  0.028067   X
    3  99.4   99.3       *      6.7  0.019283   X X
    3  99.3   99.3       *      8.6  0.019720     X X
    4  99.4   99.3       *      5.5  0.018743   X X   X
    4  99.4   99.3       *      8.7  0.019548   X X X
    5  99.4   99.3       *      6.0  0.018607   X X X X

At your request, the best subsets procedure included these variables in every model: Year2012
```

Very similar to the previous best subsets analysis, we will still choose the same model including logEle_Csp and laglogCO2. The reasons are listed below similarly:

(1) The prediction power is maximized at this model with R-sq 99.4% and adjusted R-sq 99.3% and this model is the simplest among models with same R-sq and adjusted R-sq.
(2) Although the model (Arable land, Year2012, logEle_Csp and laglogCO2) minimizes Cp, by computing the MCp Value below, the previous model has smaller MCp and the previous one is simpler.

Predictors	MCp
2 + Year2012	2.4
3 + Year2012	2.5

Therefore, the best subsets method chooses the same model. And thus the autocorrelation, unusual observations issues are all set (They have been checked as stated above). And the final model equation is re-stated as below and we are done with modeling here:

```
Regression Equation

logCO2 = -0.333 + 0.746 laglogCO2 + 0.1595 logEle_Csp - 0.0019 Year2012
```

At last, there are three more points about the regression analysis.

1 The final model seems to have addressed the autocorrelation issue given the ACF plot and runs test not statistically significant. And the diagnostics plots do not show obvious unusual observations. But the residual vs fitted value and residual vs predictor plots still reveal some structure which might be a problem although I have already done as best I can.

2 Since both the response and the predicting variables are increasing steadily with time, the leverage values tend to be higher for the earliest data in the series and the latest data in the series which makes data in this two parts more likely to be the leverage points.

3 If we are only interested in prediction, the one variable model (just include lagged variable) would be powerful enough in terms of the high R-sq.

Reference

1 World Bank Data:
http://databank.worldbank.org/data/
http://data.worldbank.org/indicator

2 Definition of the variables:

Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.

Arable land includes land defined by the FAO as land under temporary crops (double-cropped areas are counted once), temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded.

Electric power consumption measures the production of power plants and combined heat and power plants less transmission, distribution, and transformation losses and own use by heat and power plants.

Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects.