625.661 Statistical Models and Regression Course Project

DEVLOPING THE STATISTICAL MODEL FOR PREDICTING NBOCL3 CONCENTRATIONIN THE CHLORINATION OF NIOBIUM OXYCHLORIDE

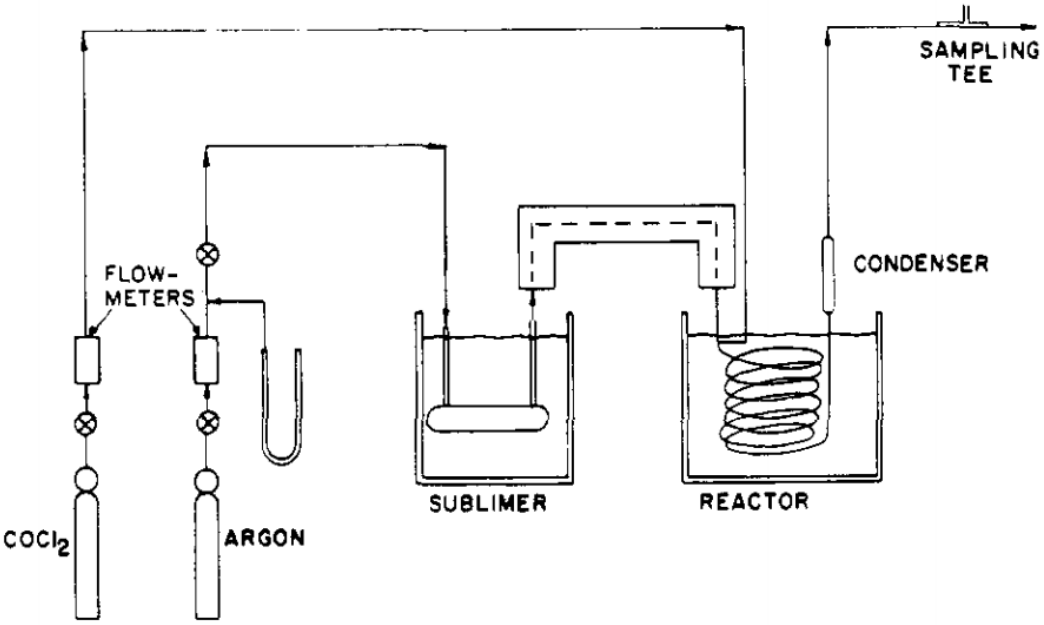


Figure 1. Tube-flow reactor system

Mohammed Momin

Table of Contents

1. Problem Description and Data

Scientists at the Institute for Atomic Research and Department of Chemical Engineering at the University of Iowa conducted an experiment to measure the effect of chlorination of Niobium Oxychloride by Phosgene in a tube-flow reactor. The following chemical equation was tested:

$$\mathrm{NbOCl_3 + COCl_2 = NbCl_5 + CO_2}$$

The independent and dependent variables were as follows:

y: $\mathrm{NbOCl_3}$ Concentration (g*mol/l)

x1: $\mathrm{COCl_2}$ Concentration (g*mol/l)

x2: Space time (sec)

x3: Molar Density (g*mol/l)

x4: Mole Fraction $\mathrm{CO_2}$

The experimental data was as follows:

Run no.	y	x1	x2	x3	x4
1	0.00045	0.0105	90.9	0.0164	0.0177
2	0.00045	0.011	84.6	0.0165	0.0172
3	0.000473	0.0106	88.9	0.0164	0.0157
4	0.000507	0.0116	488.7	0.0187	0.0082
5	0.000457	0.0121	454.4	0.0187	0.007
6	0.000452	0.0123	439.2	0.0187	0.0065
7	0.000453	0.0122	447.1	0.0186	0.0071
8	0.000426	0.0122	451.6	0.0187	0.0062
9	0.001215	0.0123	487.8	0.0192	0.0153
10	0.001256	0.0122	467.6	0.0192	0.0129
11	0.001145	0.0094	95.4	0.0163	0.0354
12	0.001085	0.01	87.1	0.0162	0.0342
13	0.001066	0.0101	82.7	0.0162	0.0323
14	0.001111	0.0099	87	0.0163	0.0337
15	0.001364	0.011	516.4	0.019	0.0161
16	0.001254	0.0117	488.4	0.0189	0.0149
17	0.001396	0.011	534.5	0.0189	0.0163

18	0.001575	0.0104	542.3	0.0189	0.0164
19	0.001615	0.0067	98.8	0.0163	0.0379
20	0.001733	0.0066	84.8	0.0162	0.036
21	0.002753	0.0044	69.6	0.0163	0.0327
22	0.003186	0.0073	436.9	0.0189	0.0263
23	0.003227	0.0078	406.3	0.0192	0.02
24	0.003469	0.0067	447.9	0.0192	0.0197
25	0.001911	0.0091	58.5	0.0164	0.0331
26	0.002588	0.0079	394.3	0.0177	0.0674
27	0.002635	0.0068	461	0.0174	0.077
28	0.002725	0.0065	469.2	0.0173	0.078

After carefully assessing the data and going through the process of statistical model development and regression analysis, we have determined the best regression model for model fitting and model prediction of $NbOCl_3$ Concentration.

2. Software used

All code and output given has come from R.

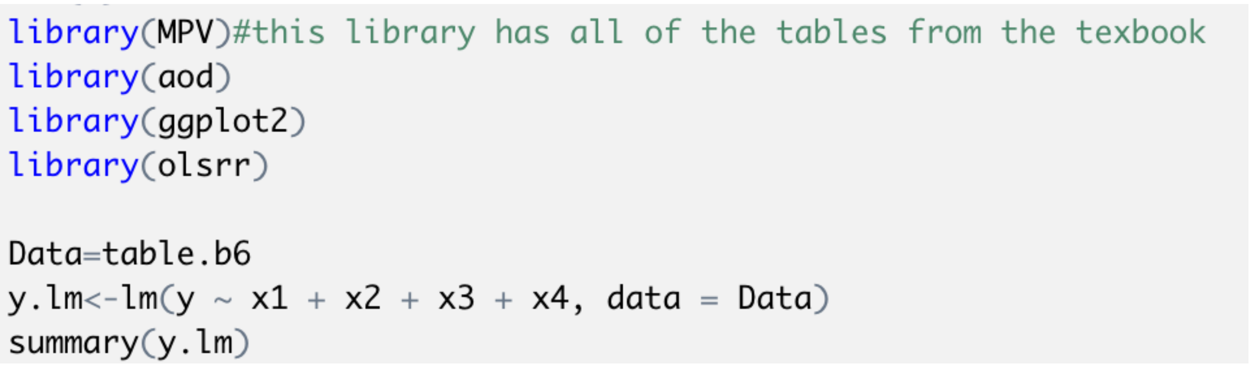
3. Assumptions

a. Multivariate normality: the residuals are normally distributed

b. No Multicollinearity: the independent variables are not highly correlated with each other.

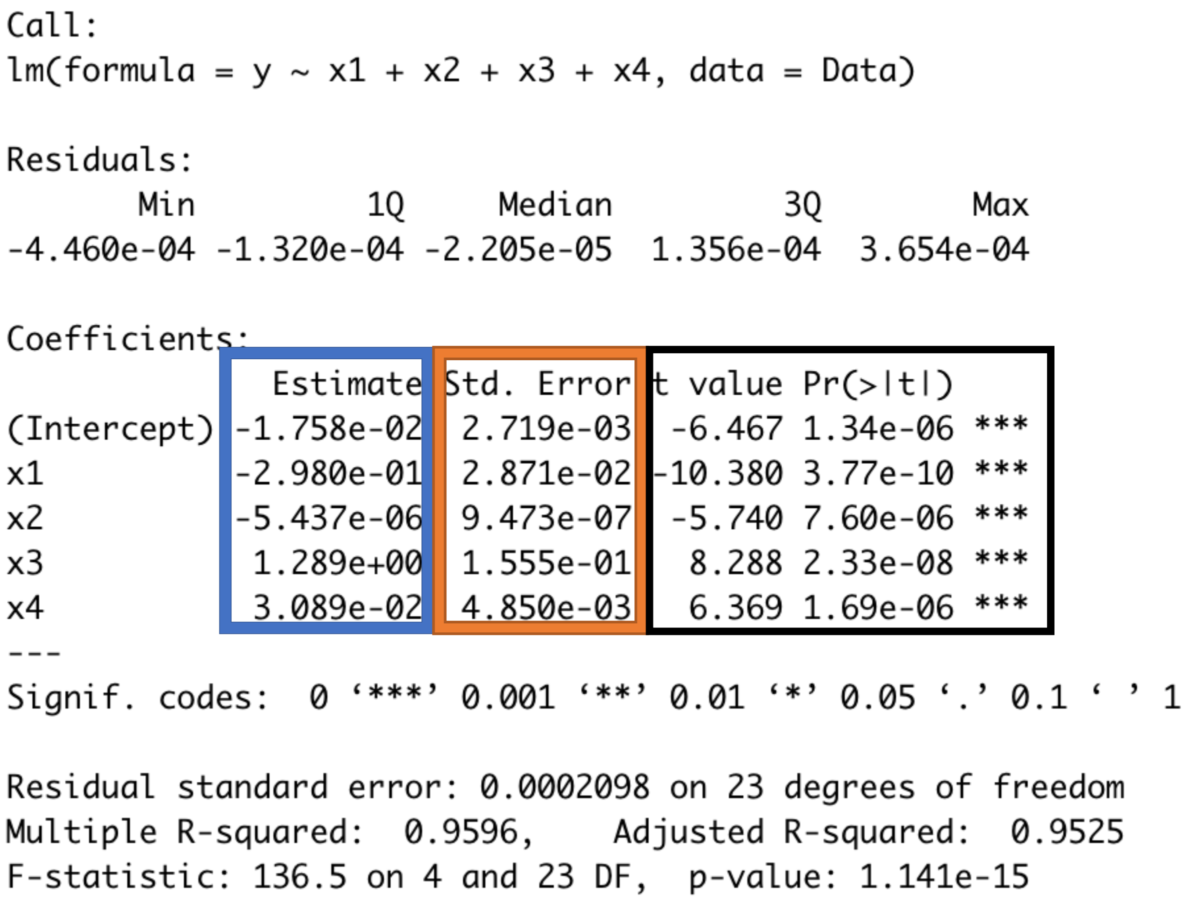
c. Homoscedasticity: all residuals have a constant finite variance

4. Model Building



```
library(MPV)#this library has all of the tables from the texbook
library(aod)
library(ggplot2)
library(olsrr)

Data=table.b6
y.lm<-lm(y ~ x1 + x2 + x3 + x4, data = Data)
summary(y.lm)
```

We start with fitting the data to the model $\underline{\boldsymbol{y}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = \beta_0 + \beta_i x_i$ (where i=1,2,3,4). The summary statistics of the multiple linear regression model given from the R code above were as follows:

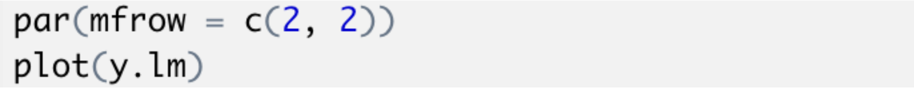


```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = Data)

Residuals:
       Min         1Q     Median         3Q        Max 
-4.460e-04 -1.320e-04 -2.205e-05  1.356e-04  3.654e-04 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.758e-02  2.719e-03  -6.467 1.34e-06 ***
x1          -2.980e-01  2.871e-02 -10.380 3.77e-10 ***
x2          -5.437e-06  9.473e-07  -5.740 7.60e-06 ***
x3           1.289e+00  1.555e-01   8.288 2.33e-08 ***
x4           3.089e-02  4.850e-03   6.369 1.69e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0002098 on 23 degrees of freedom
Multiple R-squared:  0.9596,	Adjusted R-squared:  0.9525 
F-statistic: 136.5 on 4 and 23 DF,  p-value: 1.141e-15
```

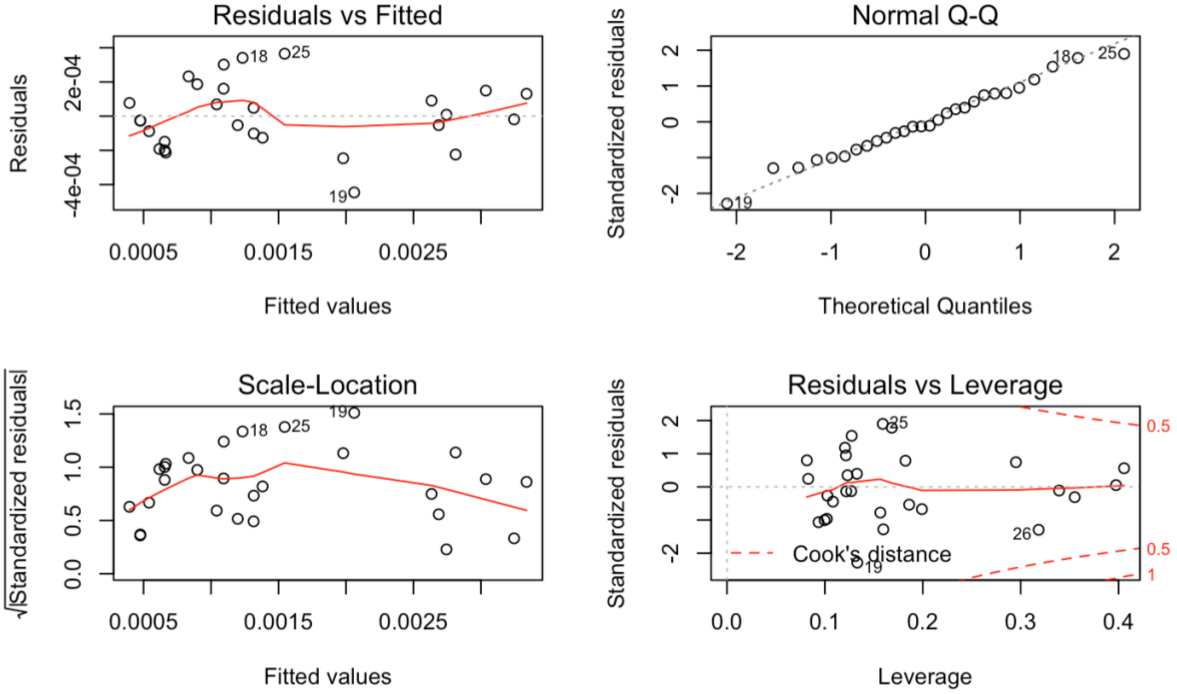
The coefficient estimates highlighted in the blue box show us the β_i values for the model $\underline{\boldsymbol{y}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$. The coefficient standard error highlighted in the orange box show us the average amount the coefficient estimates vary from the from the actual value of y. Ideally, we'd like a lower coefficient standard error relative to the estimates themselves and that is the case with this model. The t value and Pr(>|t|) will be used to assess the statistical significance of each regressor in section 7. The residual standard error, multiple R-squared, and adjusted R-squared values assess the quality of regression fit. In this case, we have an excellent set of R-squared values of around 0.95. This is interpreted as approximately 95% of the variance found in y can be explained by the x_i values. While this sounds great, further model adequacy checking will be needed because an incorrect model can yield a high R-squared value if there are certain transformations that may be needed or the relationship between some of the regressors is of a higher power.

5. Model Adequacy Checking

To further assess the adequacy of the model we must perform analysis on its residuals and outliers. The following R code was used to do check the fit of the model and make sure the assumptions that were made were not violated.



```
par(mfrow = c(2, 2))
plot(y.lm)
```



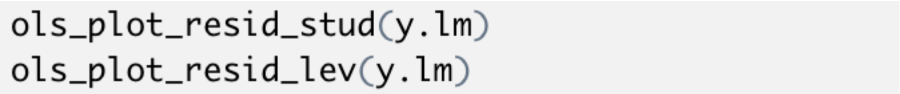
The upper left plot shows residuals (the distance between a y and the model estimate of y) and the fitted values of ***y*** which is also sometimes called y-hat. Ideally, we would like to see a smooth curve that lies close to the dotted gray line. We have this for the most part except a few points such as 18 and 25 which cause for the bump in the beginning.

The upper right plot is the qqnorm() plot of the residuals which tests the assumption of multivariate normality. Since the points lie pretty close to the dashed line, the assumption is met. If the points were exactly on the dotted lines, the residuals would be precisely normally distributed, but some deviance is expected in experimental data especially on the ends, but that deviance is very small.

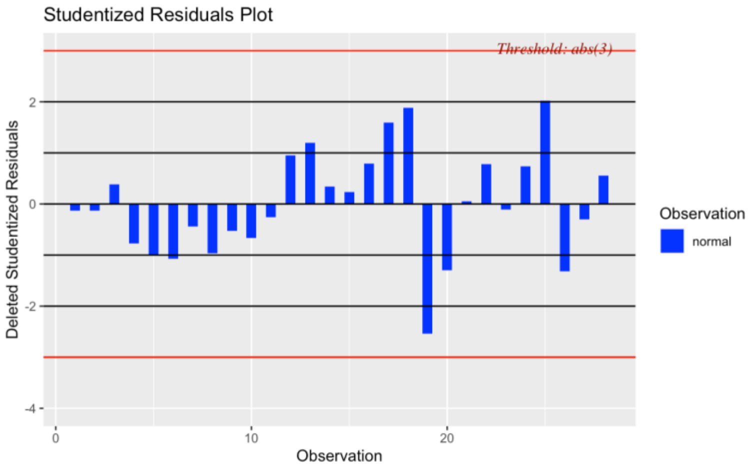
The bottom left plot shows us the square root of standardized residuals vs the fitted values. This plot tests the homoscedasticity assumption that all residuals have a finite constant variance. To satisfy this the red line should be relatively flat. Here there are a few points that prove to be outliers (points 18,19, and 25). These were the same points that were shown as outliers in the top left plot. Apart from the influence of a few outliers, the variance seems to be fairly constant to satisfy the assumption.

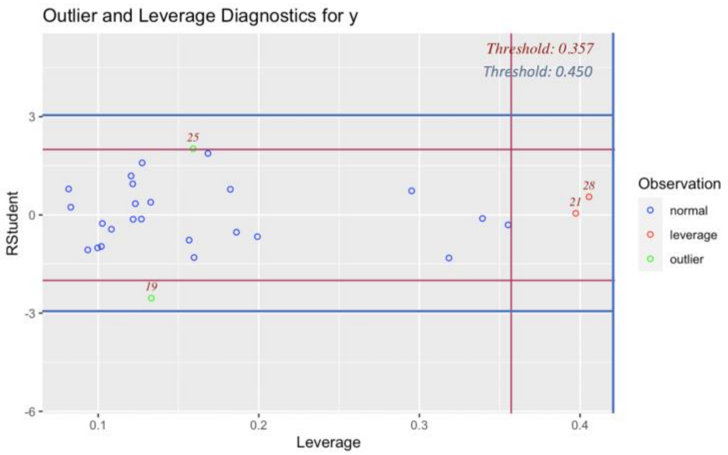
The bottom right plot shows the standardized residuals against leverage. Leverage measures how much each data point influences a regression. On this plot, you want to see that the red smoothed line stays close to the horizontal gray dashed line and that no points have a large Cook's distance (more than .05) which is true in this case.

Next we will look at Studentized Residuals and RStudent statistics with the following code:



```
ols_plot_resid_stud(y.lm)
ols_plot_resid_lev(y.lm)
```





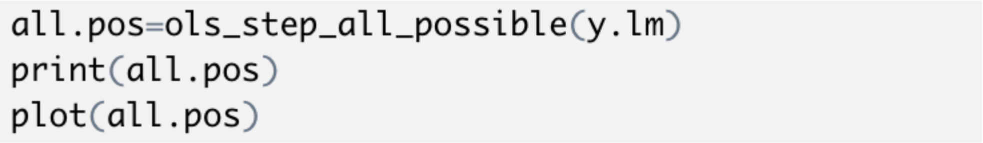
When assessing the studentized residuals plot, there were no residuals that that passed the threshold of the absolute value of 3 to be deemed outliers which is ideal. In the RStudent vs leverage plot however, there were a few points that were deemed outliers for being higher than the absolute value of 2.5. These points (19 and 25) were the same ones we saw as outliers in previous residual plots. Overall, we want a scatter plot with majority of the points having low leverage and within the threshold of RStudent which is the case in the graph above.

After performing further model adequacy checking and residual analysis, we find that all of the assumptions were met and baring a couple of outliers, the model actually fit the data very well.

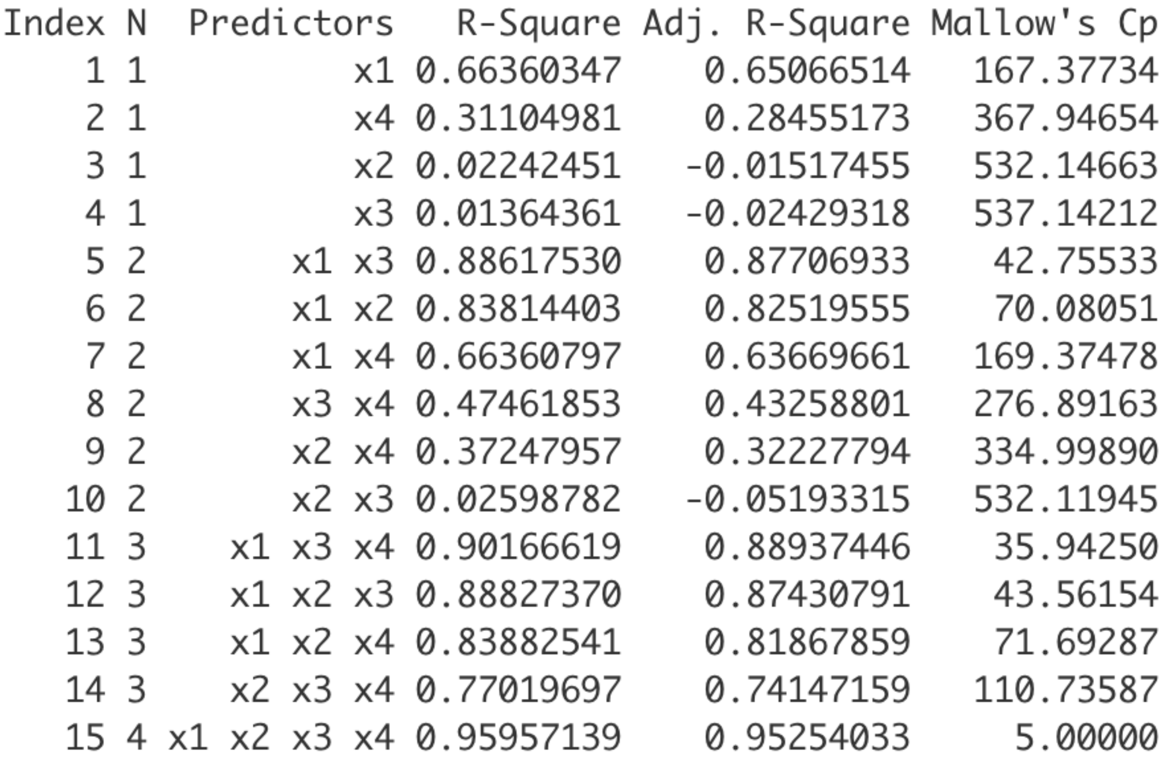
6. Variable Selection

Sometimes, the RSquared values of the model are inflated due to too many regressors some of which may be unnecessary. There are a few variable selection techniques to find the best subset of variables to include in order to find the best model: All Possible Regressions, Forward Selection, Backward Elimination, and Stepwise Regression.

The all possible regression technique assesses all possible subsets of the set of x_i's with the following code:

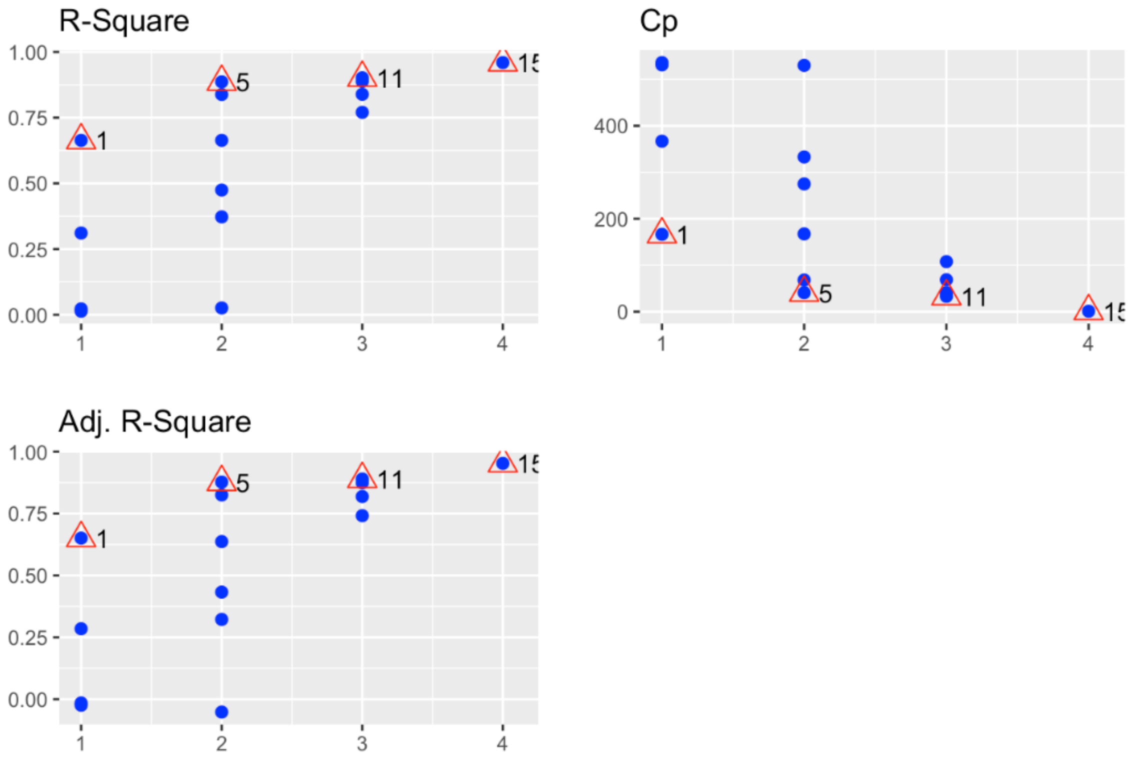


```
all.pos=ols_step_all_possible(y.lm)
print(all.pos)
plot(all.pos)
```



```
Index N   Predictors   R-Square Adj. R-Square Mallow's Cp
    1 1           x1 0.66360347    0.65066514   167.37734
    2 1           x4 0.31104981    0.28455173   367.94654
    3 1           x2 0.02242451   -0.01517455   532.14663
    4 1           x3 0.01364361   -0.02429318   537.14212
    5 2        x1 x3 0.88617530    0.87706933    42.75533
    6 2        x1 x2 0.83814403    0.82519555    70.08051
    7 2        x1 x4 0.66360797    0.63669661   169.37478
    8 2        x3 x4 0.47461853    0.43258801   276.89163
    9 2        x2 x4 0.37247957    0.32227794   334.99890
   10 2        x2 x3 0.02598782   -0.05193315   532.11945
   11 3     x1 x3 x4 0.90166619    0.88937446    35.94250
   12 3     x1 x2 x3 0.88827370    0.87430791    43.56154
   13 3     x1 x2 x4 0.83882541    0.81867859    71.69287
   14 3     x2 x3 x4 0.77019697    0.74147159   110.73587
   15 4  x1 x2 x3 x4 0.95957139    0.95254033     5.00000
```

The plots below show the output of R-Square, Adj.R-Square, and Mallow's Cp by the number of predictors n

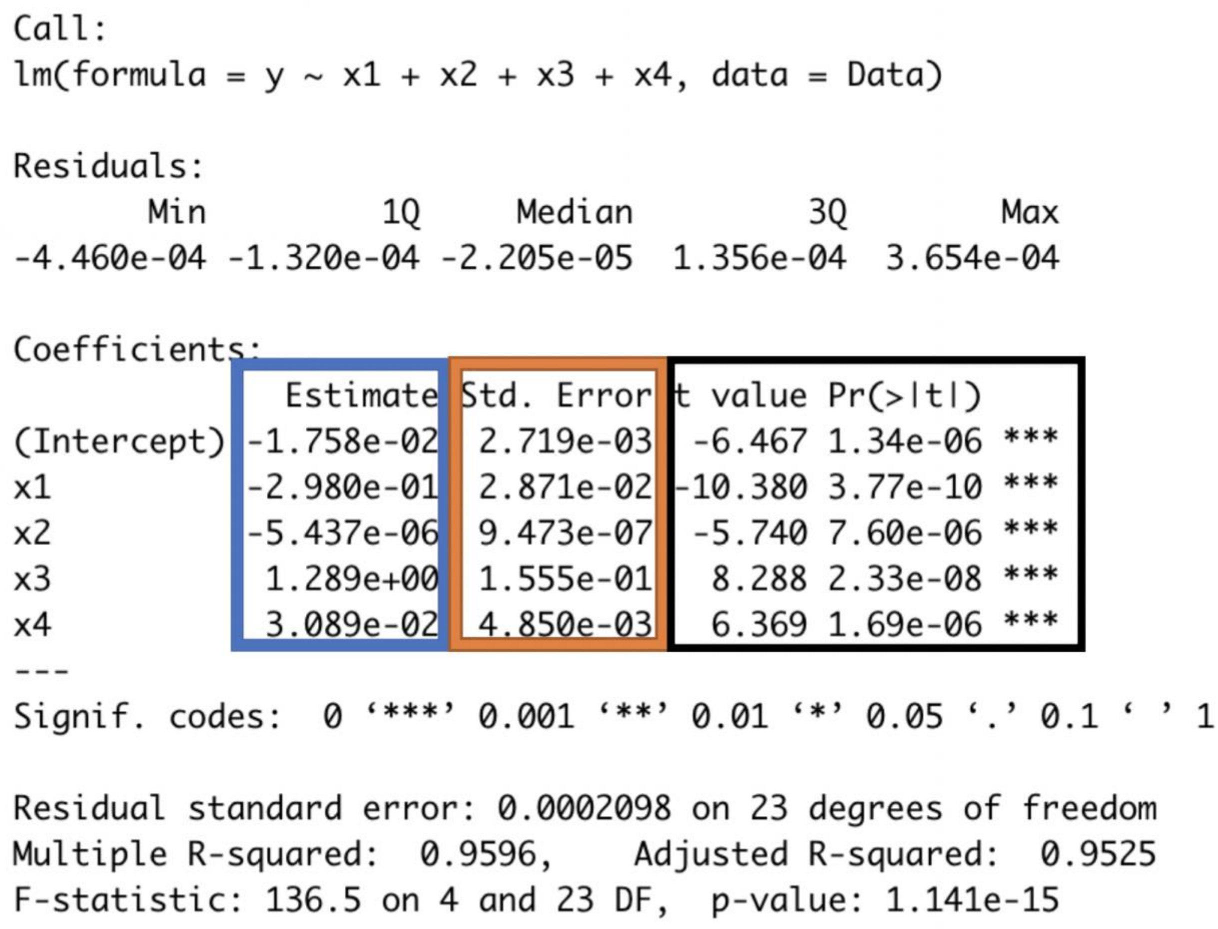


The output of the all possible regression showed that the model with all four regressors yielded the highest RSquared and RSquared Adjusted values. The Mallows Cp value is best when it is close to the number of predictors which is 4 in this case. The model with all four regressors yielded the best Cp value. This means that out of all possible subsets of the four x_i values, the model that included all four values was the best model.

Forward Selection, Backward Elimination, or Stepwise Regression are used when the All Possible Regressions method is too burdensome computationally. Since we have found the best subset using the All Possible Regressions method, there is no need to check the 3 stepwise-type procedures which would yield the same best model.

7. Statistical Significance of Each Regressor

To find the statistical significance of each regressor we have to look at the summary statistics of the model which was given in section 4 of Model Building:

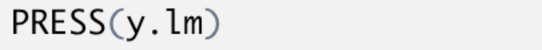


```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = Data)

Residuals:
       Min         1Q     Median         3Q        Max 
-4.460e-04 -1.320e-04 -2.205e-05  1.356e-04  3.654e-04 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.758e-02  2.719e-03  -6.467 1.34e-06 ***
x1          -2.980e-01  2.871e-02 -10.380 3.77e-10 ***
x2          -5.437e-06  9.473e-07  -5.740 7.60e-06 ***
x3           1.289e+00  1.555e-01   8.288 2.33e-08 ***
x4           3.089e-02  4.850e-03   6.369 1.69e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0002098 on 23 degrees of freedom
Multiple R-squared:  0.9596,	Adjusted R-squared:  0.9525 
F-statistic: 136.5 on 4 and 23 DF,  p-value: 1.141e-15
```

The t value in the black box shows us how many standard deviations our coefficient estimate is far away from 0. This t value is used to compute the Pr(>|t|) value which is the probability of observing any value equal or larger than the absolute value of t. The lower this value is the less unlikely it is that we are observing the relationship between the x_i values and y by chance. We typically use the cutoff point of .05 but in this case, the model coefficient P values are drastically lower than .05 which shows that each regressor is very statistically significant in the model. This allows us to reject the null hypothesis and conclude that there is a strong relationship between all of the x_i individually and y.

8. Model Validation for Prediction Purpose

R-Squared values can typically measure the quality of fit of a regression but it does not tell us how well the model can predict future values. The PRESS Statistic or the predicted residual some of squares can be used to measure the predictive power of the model. This statistic was found with the following R code:



```
PRESS(y.lm)
```

This code yielded a PRESS Statistic of 1.434039e-06. In general, the smaller the PRESS value, the better the model's predictive ability. In this case the PRESS Statistic is very small which indicates that the model has strong predictive power.

9. Conclusion

After careful statistical modeling and assessing model diagnostics and adequacy, I have determined that the best regression model for the data set is:

$\underline{\boldsymbol{y}} = - 0.01758 - 0.02980x_1 + -5.437e\text{-}06x_2 + 1.289x_3 + 0.03089x_4$

Where the variables are:

$\underline{\boldsymbol{y}}$: the predicted value of $NbOCl_3$ Concentration (g*mol/l)

x1: $COCl_2$ Concentration (g*mol/l)

x2: Space time (sec)

x3: Molar Density (g*mol/l)

x4: Mole Fraction CO_2

10. References

“ Kinetics of Chlorination of Niobium Oxychloride by Phosgene in a Tube - Flow Reactor, ” Industrial and Engineering Chemistry, Process Design Development , 11 , No. 2, 1972.

Montgomery, D. C., & Peck, E. A. (1992). Introduction to linear regression analysis. New York: Wiley.