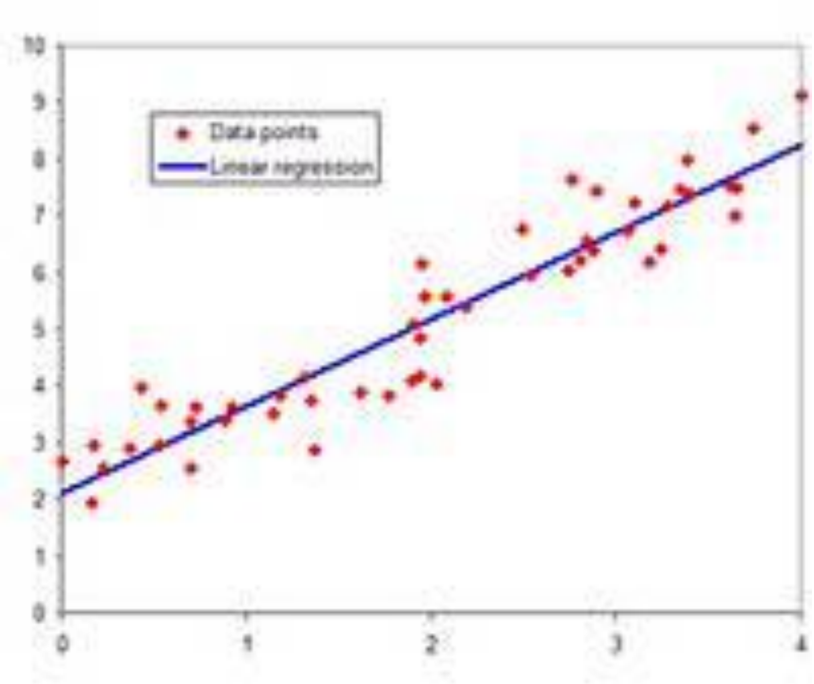


# POTATO CASE STUDY

PREDICTION OF SOLID USING MULTIPLE REGRESSION ANALAYSIS

TAIWO FAMUYIWA | MATH 7353-990-ONLINE | T00589052

TABLE OF CONTENTS

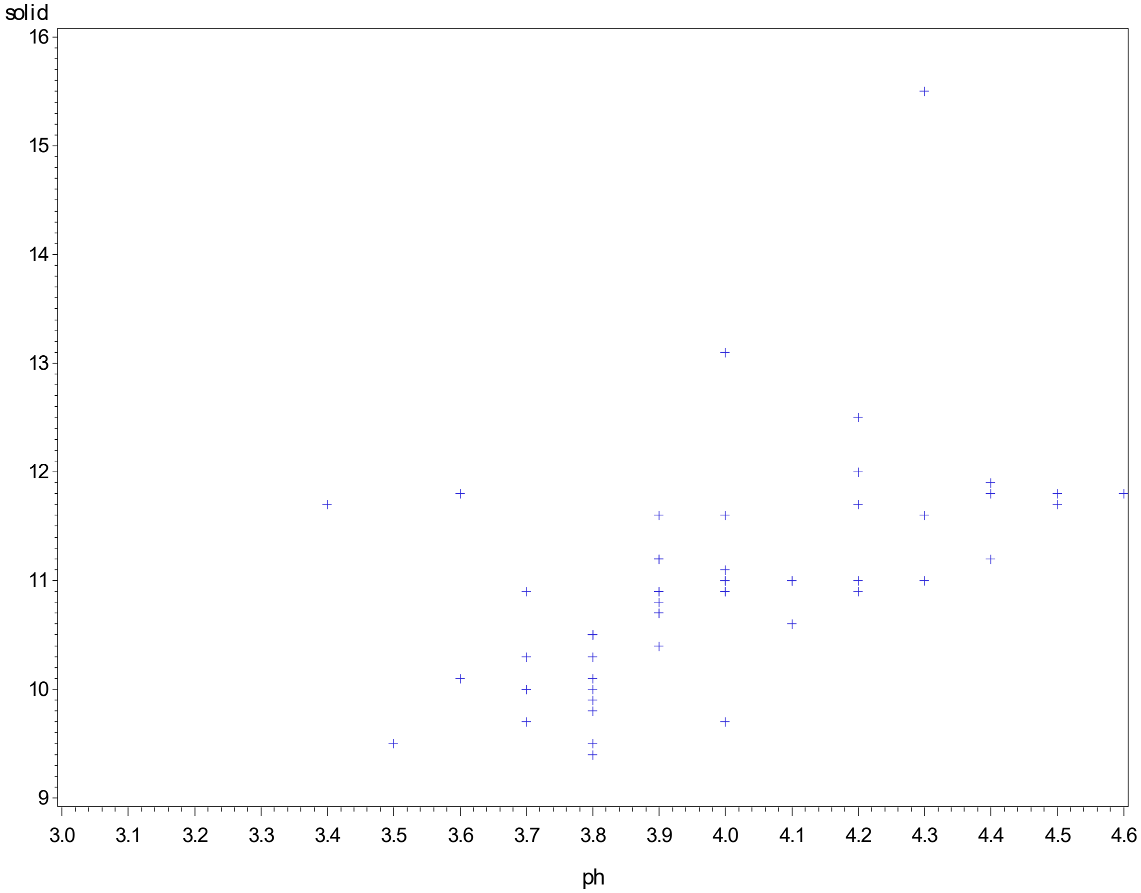
## INTRODUCTION

The project is base on the prediction of SOLID(response variable) using the predictor variables (LOWER, UPPER, THICK, VARIDRIV, AND DRUMSPD) after applying all necessary procedures for analysis.

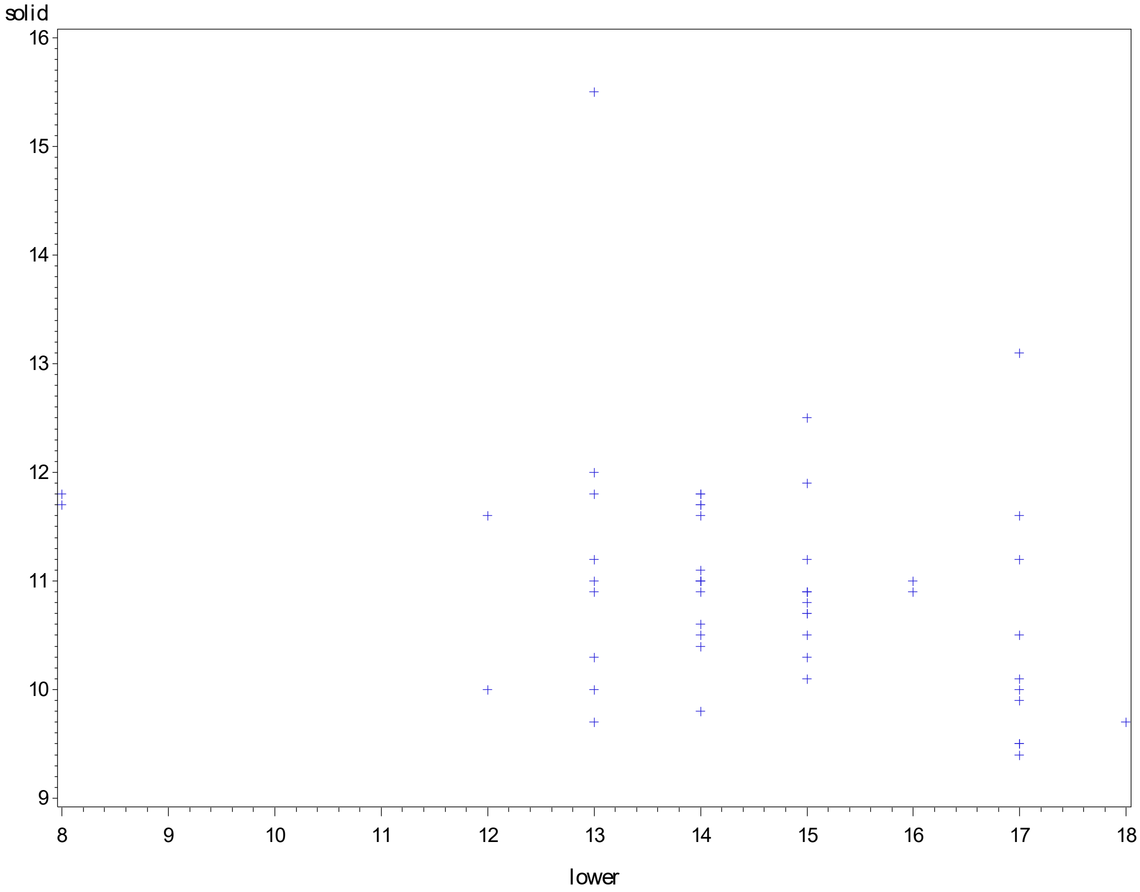
## DISCRIPTIVE STATISTICS

## SCATTER PLOT

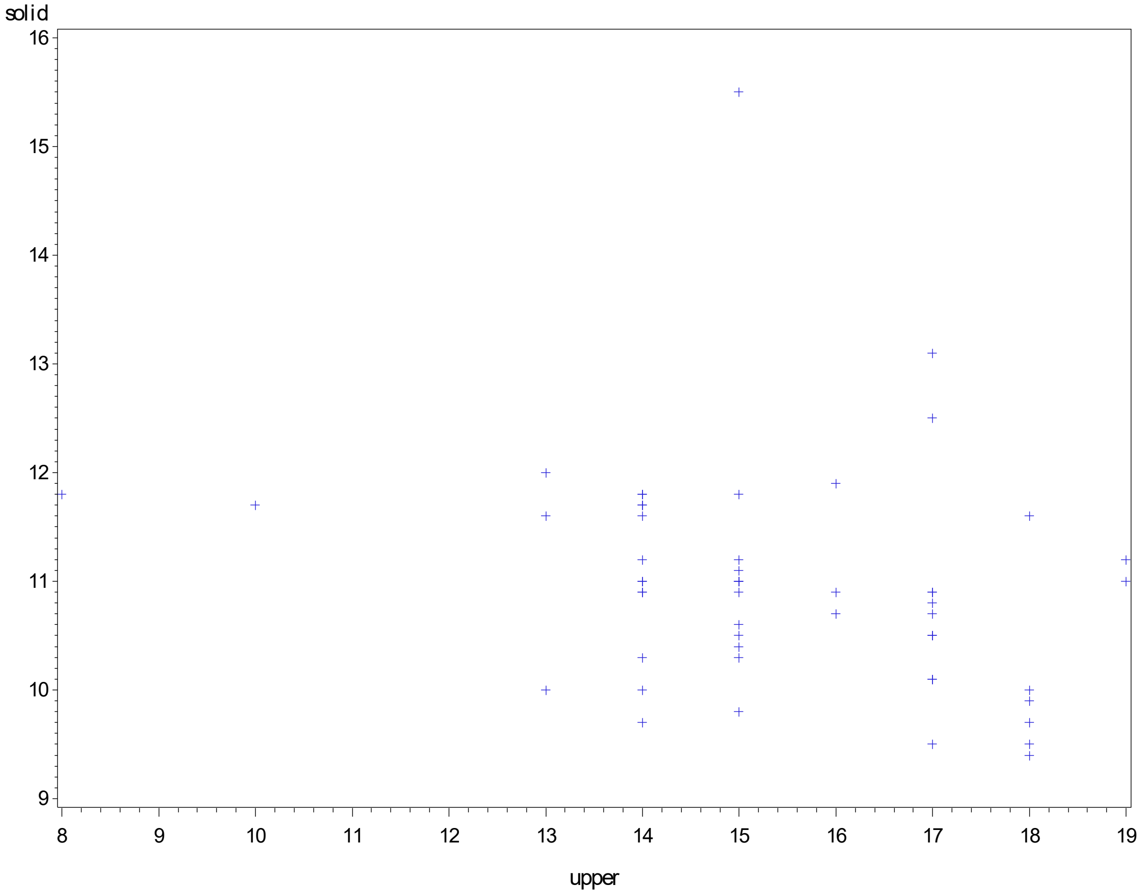
*In order to explain the relationship between Solid(response variable) and predictors variables (PH, Lower, Upper, Thick, Varidriv, and Drumspd), apply scatter plot to show this correlation. The following scatter plots show the relationship between response variable(Solid) and predictor variables.*



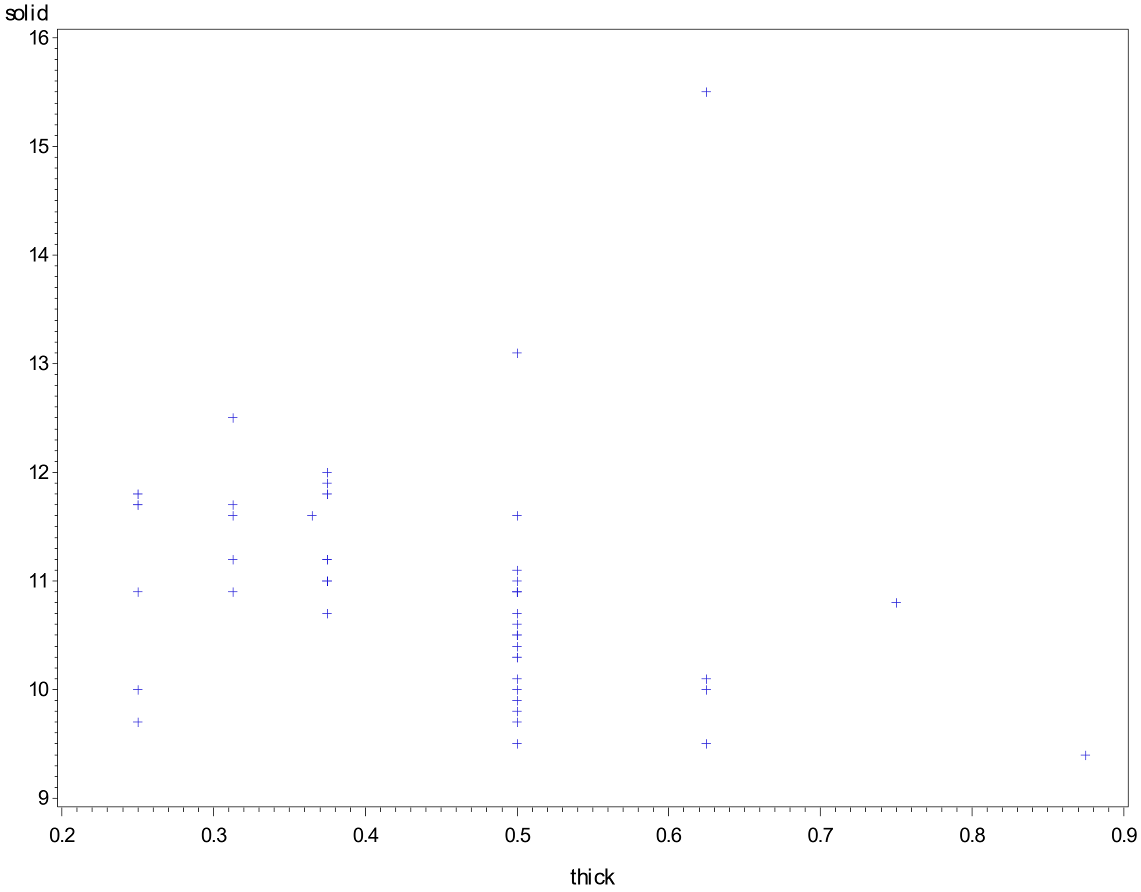
*A moderate positive correlation exist between between solid and PH*



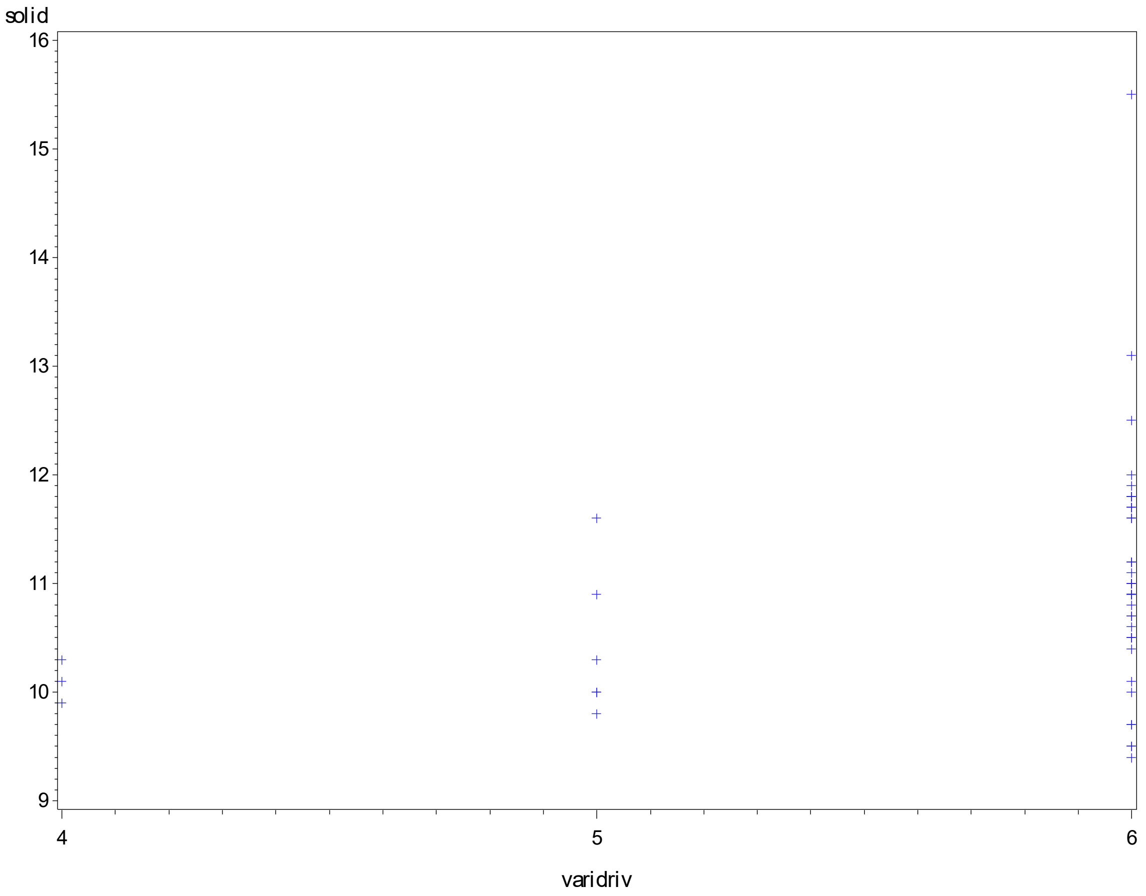
***There is little or no correlation between the Sollid and Lower.***



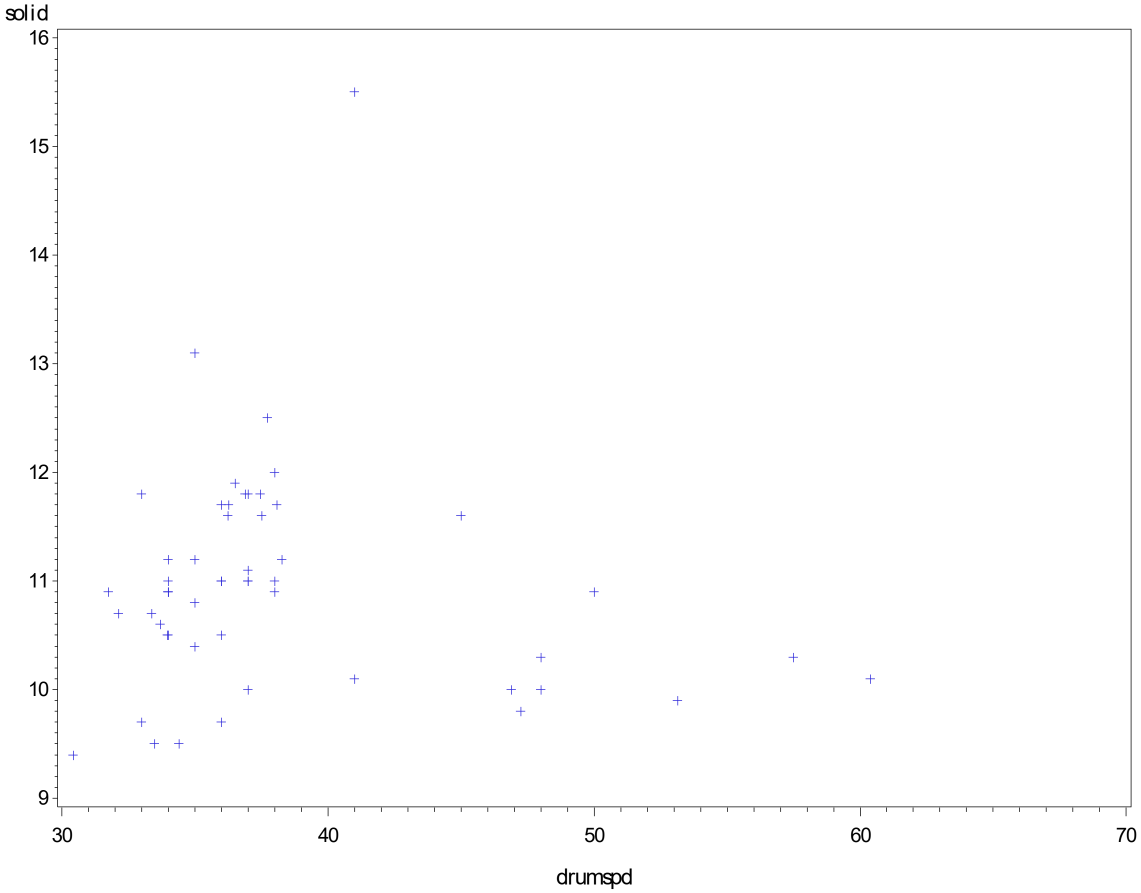
***There is little or no relationship between solid or Upper.***



***There is little or no relationship between Solid and Thick.***



***Their is little or no relationship between Solid and varidriv***



*The plot above shows a moderate negative correlation between  Solid and Drumspd.*

## CORRELATION COEFIENT

***To investigate this further, determine the correlation coefficient between Solid and predictive variables using pearson correlation of Proc Corr (SAS). There exist the following results***

| Pearson Correlation Coefficients<br>Prob > \|r\| under H0: Rho=0<br>Number of Observations | |
|---|---|
| | **solid** |
| **ph** | 0.53970<br><.0001<br>53 |
| **lower** | -0.30843<br>0.0246<br>53 |
| **upper** | -0.26689<br>0.0534<br>53 |
| **thick** | -0.26553<br>0.0546<br>53 |
| **varidriv** | 0.29176<br>0.0340<br>53 |
| **drumspd** | -0.13078<br>0.3506<br>53 |

***The correlation between solid and ph is moderate positive relationship with $p-value$ less than 0.05. Which means the correlation is signfcant.***

***The correlation between solid and lower, upper, thick and drumspdare weak negative linear relationship. The correlation between Solid and lower is significant at $p - value$ of $0.0246$***

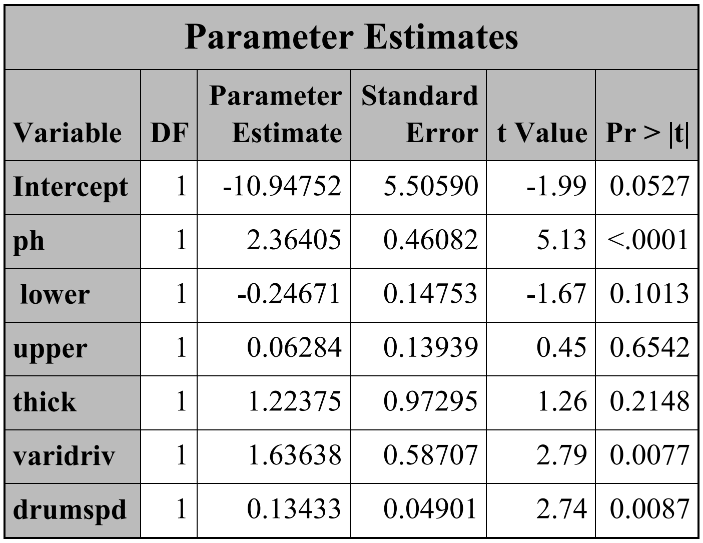
***The correlation between Solid and Varidriv is weak positive linear linear relationship but with a it is signifanct at $0.0340 < 0.05$***

## STATISTICAL ANALYSIS

*It is important to note that correlation does not mean causation.Thus, further regression analysis should be carried out.*

*Because there exist two or more predictor variables, multiple linear regression will be apply to fit this model.*

*The result when all predictor's variables are consider in fitting the model is as follows :*



**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **Intercept** | 1 | -10.94752 | 5.50590 | -1.99 | 0.0527 |
| **ph** | 1 | 2.36405 | 0.46082 | 5.13 | <.0001 |
| **lower** | 1 | -0.24671 | 0.14753 | -1.67 | 0.1013 |
| **upper** | 1 | 0.06284 | 0.13939 | 0.45 | 0.6542 |
| **thick** | 1 | 1.22375 | 0.97295 | 1.26 | 0.2148 |
| **varidriv** | 1 | 1.63638 | 0.58707 | 2.79 | 0.0077 |
| **drumspd** | 1 | 0.13433 | 0.04901 | 2.74 | 0.0087 |

$Solid = -10.948 + 2.3641ph - 0.2467lower + 0.0628upper +$
$1.2238thick + 1.6364varidriv + 0.1343drumspd$
$where\ \hat{\beta}_0 = -10.948, \hat{\beta}_1 = 2.3641, \hat{\beta}_2 = -0.2467, \hat{\beta}_3 = 0.0628,$
$\hat{\beta}_4 = 1.2238, \hat{\beta}_5 = 1.6364, \hat{\beta}_6 = 0.1343\ \ are\ least\ square\ estimates.$
$or\ coefficient$

### *Interpretation of the least square estimates*

*Holding all other regressors constant, if lower increases, solid will reduce by* 0.2467. *For* $\hat{\beta}_3 = 0.0628$ *means that, holding all other regressors constant, as upper increases so will solid increase in value by* 0.0628. *for* $\hat{\beta}_1 = 2.3641$, *holding all other regressor fixed, an increase in ph will lead to an increase in solid by* 2.3641.

## ***Statistical Inferences on $\beta's$***

*In order to determine which predictor varibles have statistically significant effects on the response variable, hypothesis testing is carried out.*

*i.e* $H_0$: $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = \hat{\beta}_6 = 0$ *Vs* $H_1$ : *At least one* $\beta_i \neq 0$

*reject* $H_0$ *if* $p - value < \alpha$

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 29.65852 | 4.94309 | 9.29 | <.0001 |
| **Error** | 46 | 24.48035 | 0.53218 | | |
| **Corrected Total** | 52 | 54.13887 | | | |

*Thus, from the table, the* $p - value = 0.0001 < \alpha = 0.05$, *reject* $H_0$.
*It means there is at least one* $\beta_i$ *that is significant.*
*To test the hypothesis* $H_0$: $\hat{\beta}_1 = 0$, $p - value$ $\hat{\beta}_1 = 0.0001 < \alpha = 0.05$.
*Reject* $H_0$, *therefore, Ph may remain in the model.*

*Also for* $H_0$: $\hat{\beta}_1 = 0$ , *the* $p - value = 0.1013 > 0.05$, *fail to reject* $H_0$. *Lower may not be in the model. Do this for other* $\hat{\beta}_{i,}$ *it is obvious that* $x_1$ *is the only regressor that seems to have a greater likelihood of remaining in the model. This shows that further analysis can still be done.*

| | | | |
|---|---|---|---|
| **Root MSE** | 0.72951 | **R-Square** | 0.5478 |
| **Dependent Mean** | 10.96604 | **Adj R-Sq** | 0.4888 |
| **Coeff Var** | 6.65243 | | |

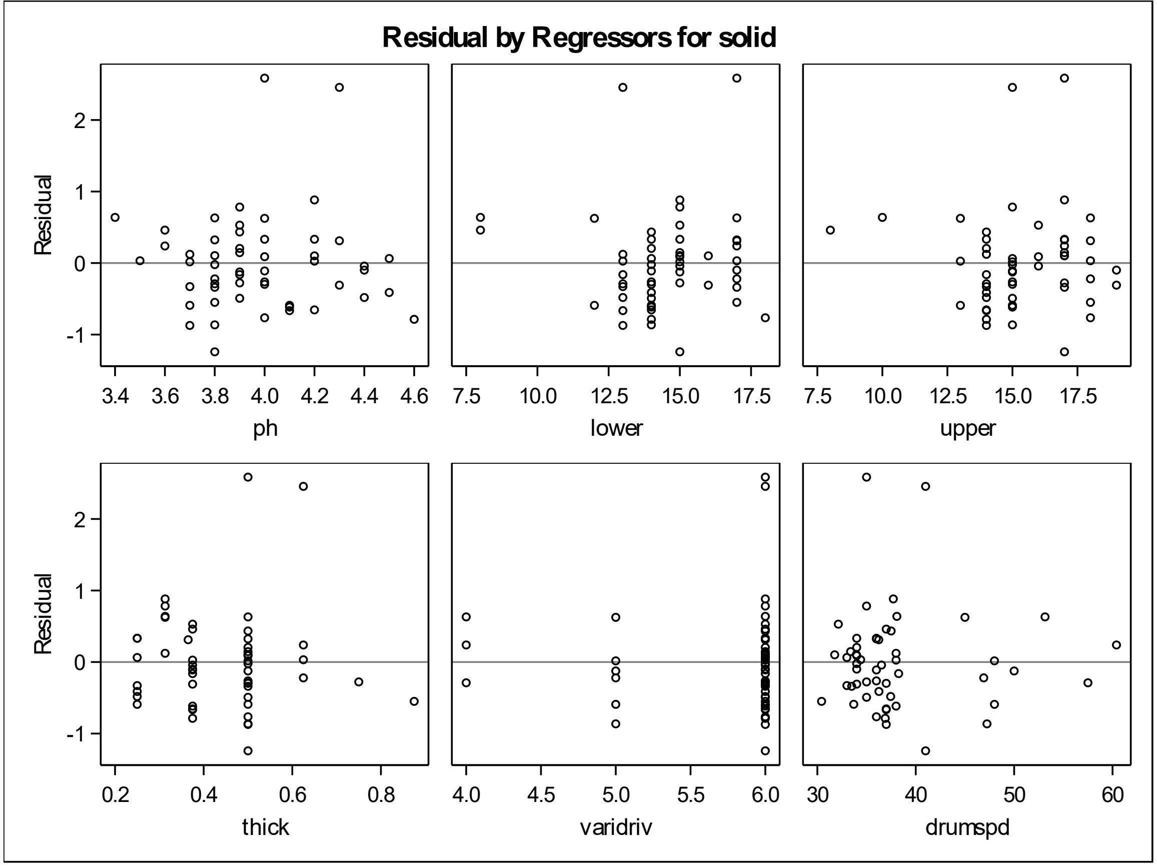
*Also, the coefficient of determination is* $R^2 = 0.5478$. *Which means the overall fit of the model is not too good. About* 54.78% *of the total variation is explained by multiple regression model. To investigate this difficulty, a model checking procedure must be applied.*

## *MODEL CHECKING (DIAGNOSTIC TESTING)*

*The following test will be carried out*:

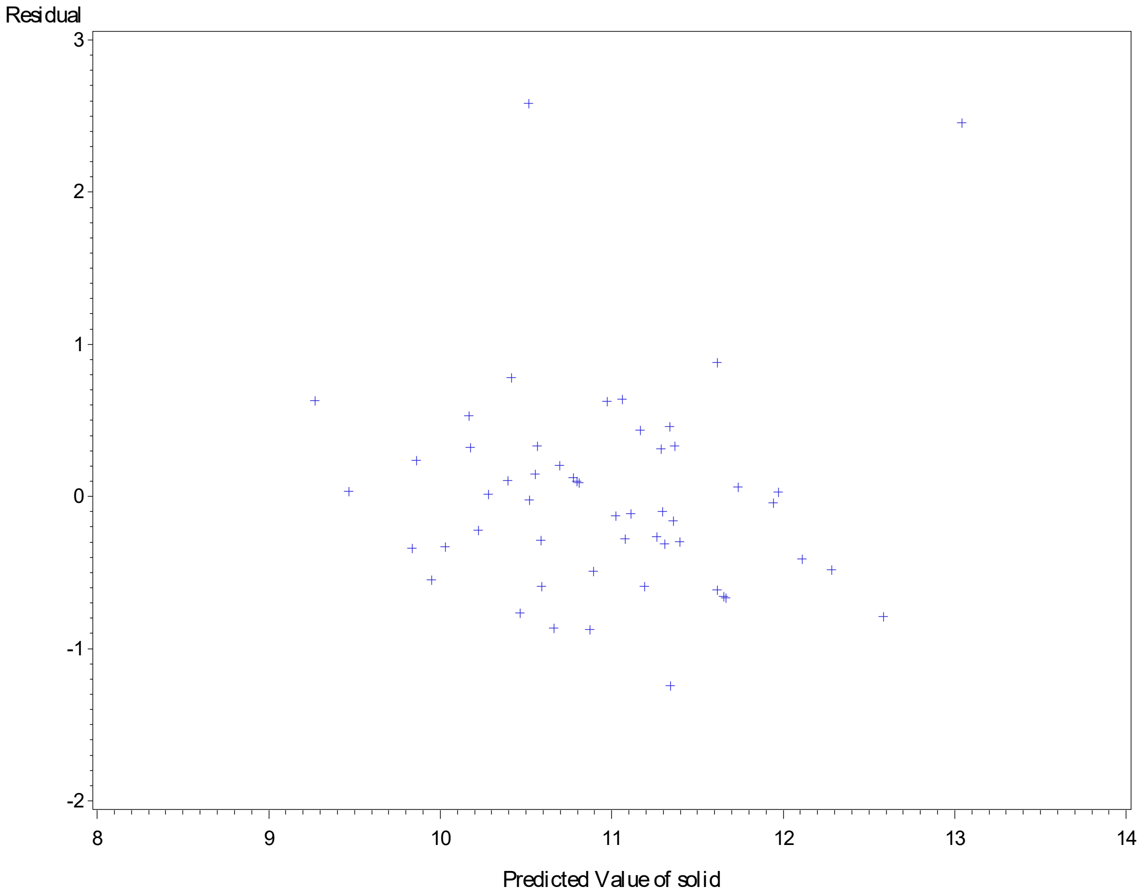
- Linearity
- Constant variance for the errors
- Normally distributed error
- Test for outliers and influential observation.

### Linearity



*The main idea of linearity is to check if the regression of solid on EACH regressor is linear. From the graph above, the residual plot against ph and drumspd fairly exhibit random scatter around zero. All other plot show systematic pattern that indicates a corresponding deviation form linearity. To deduce from this, the regression of y on ph and drumspd may be fairly linear.*

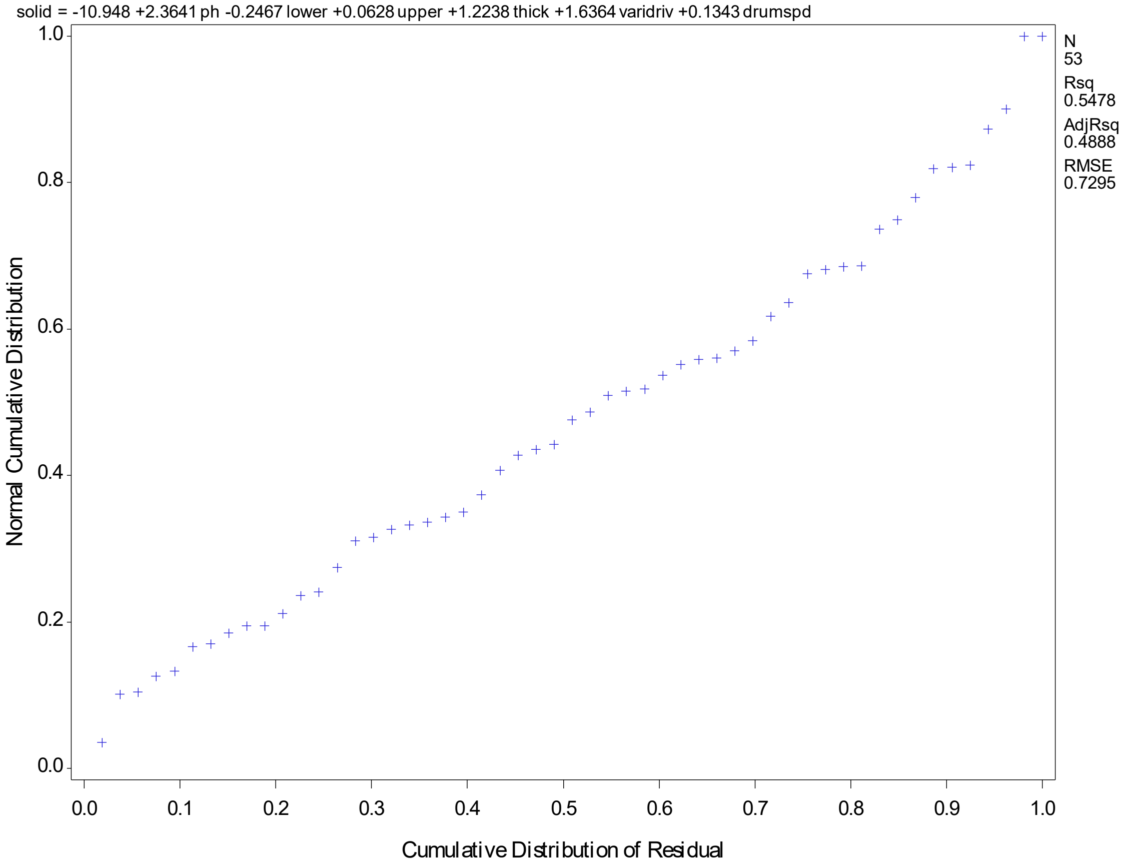
## CHECKING FOR CONSTANT VARIANCE



*Using proc reg of SAS can be used to determine the constant variance. From above diagram, it can be seen that the residuals (error term) take on negative values with small fitted values than the positive value. It can be seen that the residual is fairly scatter around zero. Less of constant variance was esterblish.*

## CHECKING FOR NORMALITY

*Another test that should be carried out is checking for normality of the residual.*



*This plot seems to be close to normality because it is nearly normal.*
*This suggests that the error distribution is normal or agreement with normality.*

## EFFECT OF INDIVIDUAL CASES

Most times, outliers might affect the outcome of a model. If discovered, it can be removed and the model can be re-fitted.
To get the best model to predict Solid, the next aspect is to check for outliers.

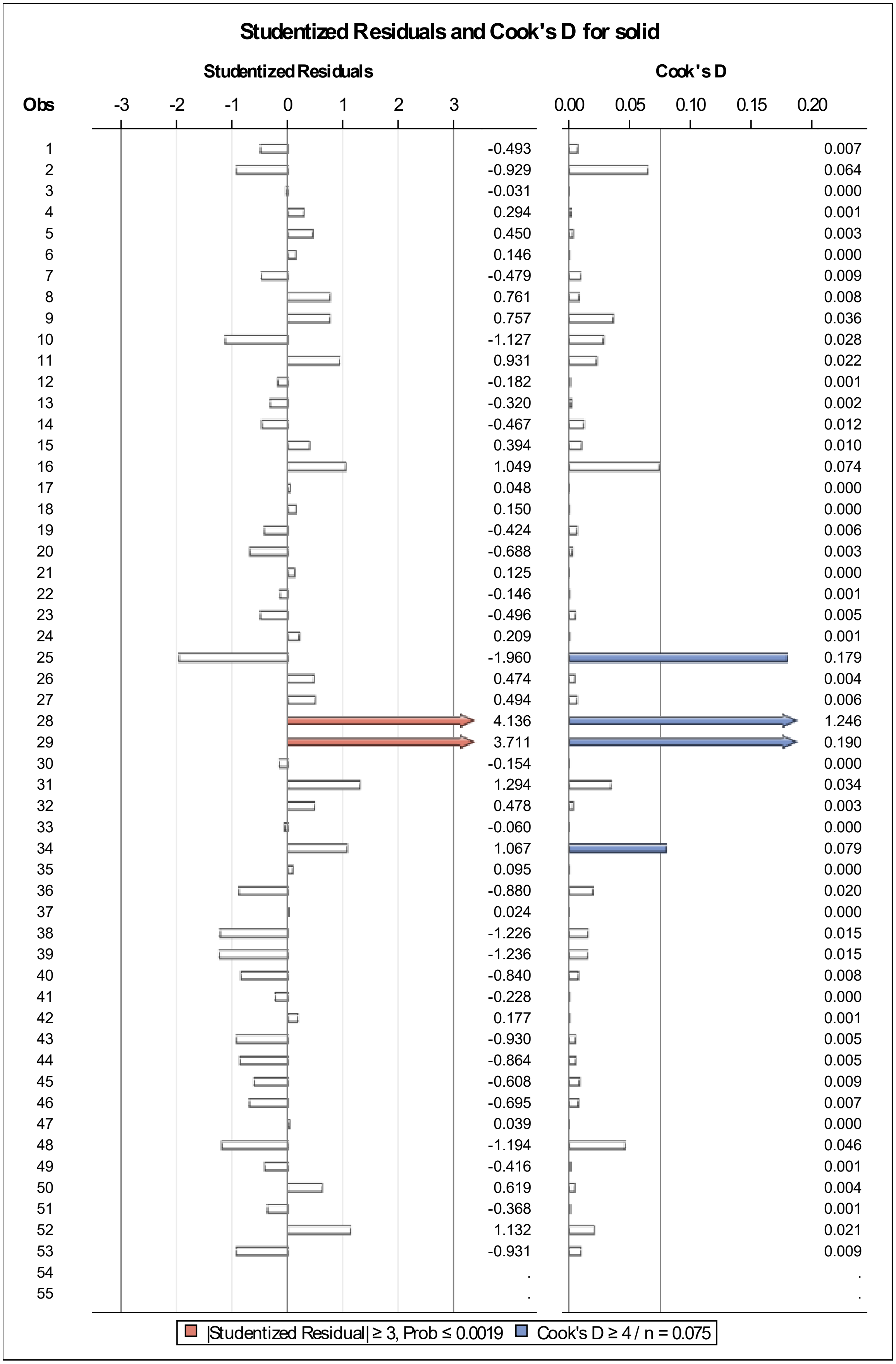
## CHECKING FOR OUTLIERS AND INFLUENTIAL

*The dot plot of the studentized residuals below shows that the studendized value for observation* 28 *and* 29 *is high i.e* $4.136 > 2.5$ *and* $3.711 > 2.5$.

## CHECING FOR INFLUENTIAL

*The Cook's distance measures the influence of each individual case has on a given statistical procedure if the conclusion of the analysis are significantly altered when the case is omitted from the analysis. Taking a look at the observation below, the following observations have a high Cook's distance; observation* 25 (*cook's distance* $= 0.179$), *observation* 28 (*Cook's distance* $= 1.246$), *observation*29 (*Cook's distance* $= 0.190$) *and observation* 34 (*Cook's distance* $= 0.190$) .

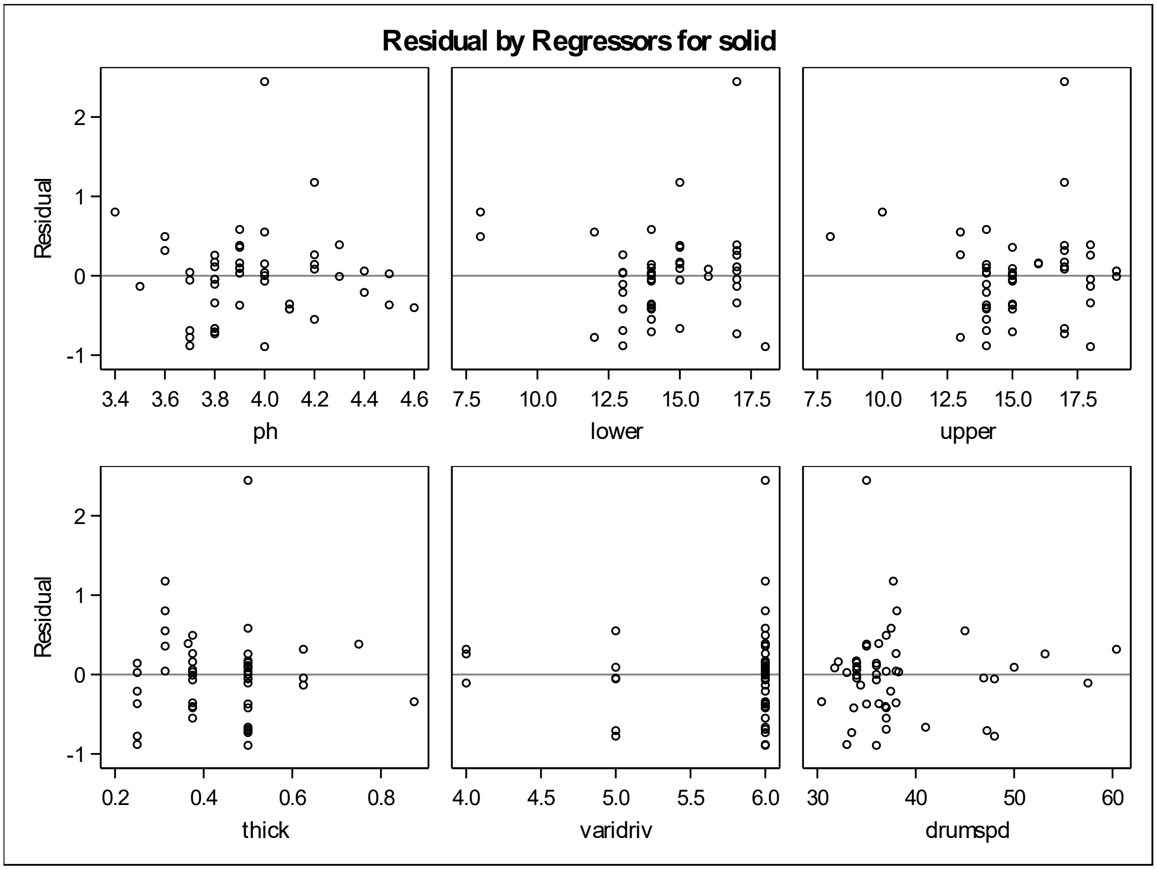
The next thing to do is to remove observation 25, 28, 29 and 34 and refit the model to show if there will be a significant difference from the first model.


Studentized Residuals and Cook's D for solid
Studentized Residuals
Cook's D
Obs
-3
-2
-1
0
1
2
3
0.00
0.05
0.10
0.15
0.20
1 -0.493 0.007
2 -0.929 0.064
3 -0.031 0.000
4 0.294 0.001
5 0.450 0.003
6 0.146 0.000
7 -0.479 0.009
8 0.761 0.008
9 0.757 0.036
10 -1.127 0.028
11 0.931 0.022
12 -0.182 0.001
13 -0.320 0.002
14 -0.467 0.012
15 0.394 0.010
16 1.049 0.074
17 0.048 0.000
18 0.150 0.000
19 -0.424 0.006
20 -0.688 0.003
21 0.125 0.000
22 -0.146 0.001
23 -0.496 0.005
24 0.209 0.001
25 -1.960 0.179
26 0.474 0.004
27 0.494 0.006
28 4.136 1.246
29 3.711 0.190
30 -0.154 0.000
31 1.294 0.034
32 0.478 0.003
33 -0.060 0.000
34 1.067 0.079
35 0.095 0.000
36 -0.880 0.020
37 0.024 0.000
38 -1.226 0.015
39 -1.236 0.015
40 -0.840 0.008
41 -0.228 0.000
42 0.177 0.001
43 -0.930 0.005
44 -0.864 0.005
45 -0.608 0.009
46 -0.695 0.007
47 0.039 0.000
48 -1.194 0.046
49 -0.416 0.001
50 0.619 0.004
51 -0.368 0.001
52 1.132 0.021
53 -0.931 0.009
54 . .
55 . .
🟥 |Studentized Residual| ≥ 3, Prob ≤ 0.0019
🟦 Cook's D ≥ 4 / n = 0.075

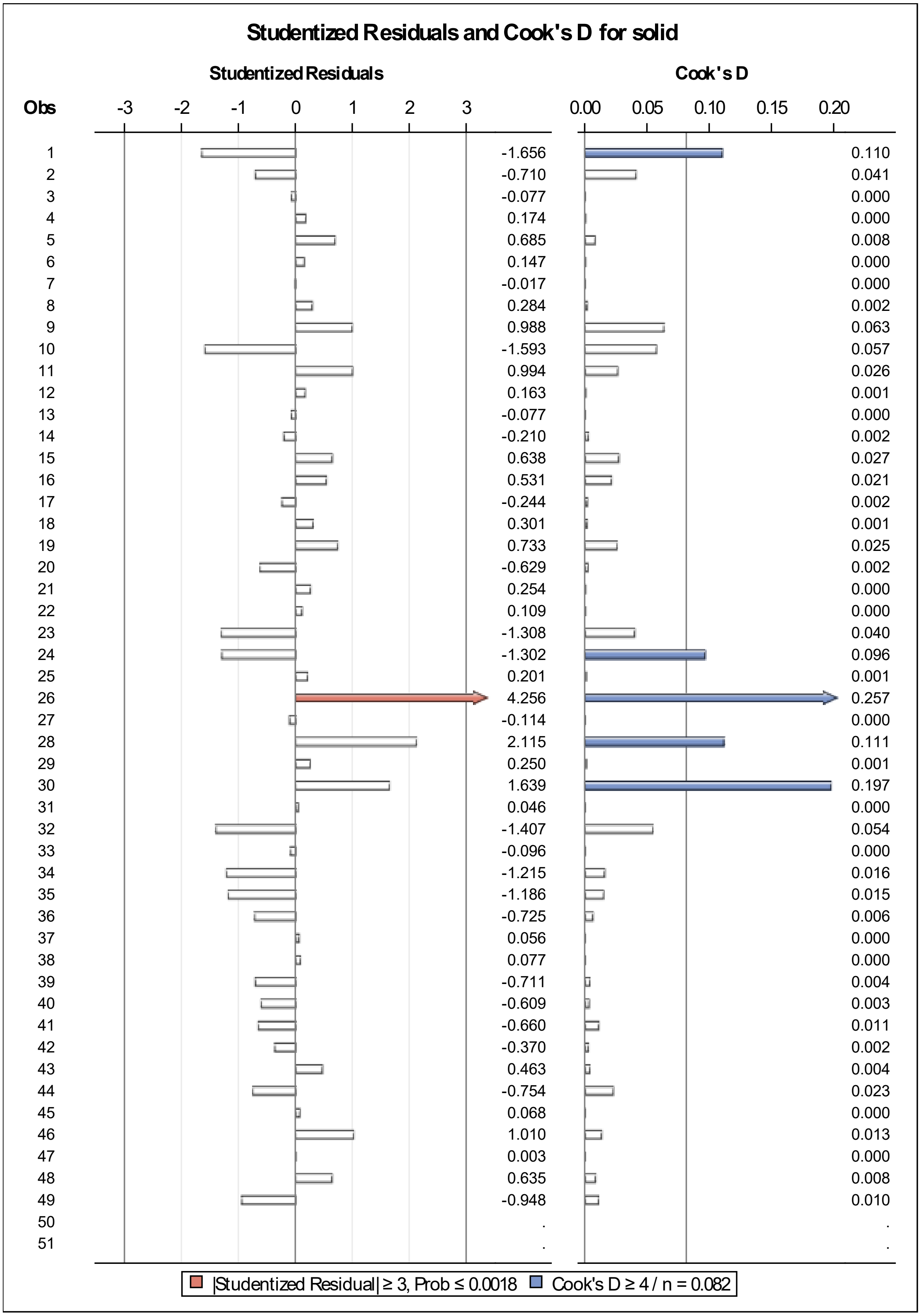
*After the refit of the model, the new regression equation is*

$$y = -0.866 + 1.6466ph - 0.0336lower - 0.0905upper - 0.5577thick$$
$$+0.8589varidriv + 0.0652drumpspd$$

*Their seems to be a significant different between the first model and the second model. All except the coefficent of ph is positive for the first model. But for the second model, most of the coefficicient of each regressor is negative. For example, an increase in upper leads to an reduce in solid. Thus, the removal of the outliers and influencial have significant difference on the new model.*



*The residual plot shown above shows a little difference from original.* *Observation* 26 *with* $d_i = 4.256$ *will not make musch different.*


Studentized Residuals and Cook's D for solid
Studentized Residuals
Cook's D
Obs
-3
-2
-1
0
1
2
3
0.00
0.05
0.10
0.15
0.20
1 -1.656 0.110
2 -0.710 0.041
3 -0.077 0.000
4 0.174 0.000
5 0.685 0.008
6 0.147 0.000
7 -0.017 0.000
8 0.284 0.002
9 0.988 0.063
10 -1.593 0.057
11 0.994 0.026
12 0.163 0.001
13 -0.077 0.000
14 -0.210 0.002
15 0.638 0.027
16 0.531 0.021
17 -0.244 0.002
18 0.301 0.001
19 0.733 0.025
20 -0.629 0.002
21 0.254 0.000
22 0.109 0.000
23 -1.308 0.040
24 -1.302 0.096
25 0.201 0.001
26 4.256 0.257
27 -0.114 0.000
28 2.115 0.111
29 0.250 0.001
30 1.639 0.197
31 0.046 0.000
32 -1.407 0.054
33 -0.096 0.000
34 -1.215 0.016
35 -1.186 0.015
36 -0.725 0.006
37 0.056 0.000
38 0.077 0.000
39 -0.711 0.004
40 -0.609 0.003
41 -0.660 0.011
42 -0.370 0.002
43 0.463 0.004
44 -0.754 0.023
45 0.068 0.000
46 1.010 0.013
47 0.003 0.000
48 0.635 0.008
49 -0.948 0.010
50 . .
51 . .
🟥 |Studentized Residual| ≥ 3, Prob ≤ 0.0018
🟦 Cook's D ≥ 4 / n = 0.082

## MODEL SELECTION

Having talked about descriptive statistic, diagnostic testing, the next point of call is  model selection. Under model selection, we want to know which of the models is best predictors of solid.

*To do this, apply cp critirion. The main goal is to identify the  subset of X for which Cp value is small, Cp value is near p.*

| Number in Model | C(p) | R-Square | Variables in Model |
|---:|---:|---:|:---|
| **4** | 4.9036 | 0.5291 | ph lower varidriv drumspd |
| **5** | 5.2032 | 0.5458 | ph lower thick varidriv drumspd |
| **5** | 6.5820 | 0.5323 | ph lower upper varidriv drumspd |
| **4** | 6.8581 | 0.5099 | ph upper varidriv drumspd |
| **6** | 7.0000 | 0.5478 | ph lower upper thick varidriv drumspd |
| **5** | 7.7964 | 0.5203 | ph upper thick varidriv drumspd |
| **3** | 8.9425 | 0.4697 | ph lower thick |
| **2** | 9.1183 | 0.4484 | ph lower |
| **3** | 9.4205 | 0.4650 | ph varidriv drumspd |
| **4** | 10.5317 | 0.4738 | ph lower thick varidriv |
| **3** | 10.6954 | 0.4525 | ph lower varidriv |
| **4** | 10.8407 | 0.4707 | ph lower thick drumspd |
| **4** | 10.8969 | 0.4702 | ph lower upper thick |
| **3** | 11.0021 | 0.4495 | ph lower drumspd |
| **3** | 11.0055 | 0.4495 | ph lower upper |
| **4** | 11.3922 | 0.4653 | ph thick varidriv drumspd |
| **2** | 12.0814 | 0.4192 | ph upper |

| Number in Model | C(p) | R-Square | Variables in Model |
|---:|---:|---:|:---|
| **5** | 12.5137 | 0.4740 | ph lower upper thick varidriv |
| **4** | 12.6299 | 0.4532 | ph lower upper varidriv |
| **3** | 12.6697 | 0.4331 | ph upper thick |
| **5** | 12.7694 | 0.4715 | ph lower upper thick drumspd |
| **4** | 12.8472 | 0.4510 | ph lower upper drumspd |
| **3** | 13.2591 | 0.4273 | ph upper varidriv |
| **4** | 13.7837 | 0.4418 | ph upper thick varidriv |
| **3** | 14.0641 | 0.4194 | ph upper drumspd |
| **4** | 14.6643 | 0.4332 | ph upper thick drumspd |
| **1** | 23.0987 | 0.2913 | ph |
| **2** | 23.5037 | 0.3070 | ph varidriv |
| **2** | 24.4084 | 0.2981 | ph thick |
| **3** | 24.9601 | 0.3123 | ph thick varidriv |
| **2** | 25.0113 | 0.2921 | ph drumspd |
| **3** | 26.3116 | 0.2990 | ph thick drumspd |
| **3** | 28.6278 | 0.2762 | thick varidriv drumspd |
| **3** | 29.0840 | 0.2718 | lower varidriv drumspd |

| Number in Model | C(p) | R-Square | Variables in Model |
|---:|---:|---:|:---|
| **2** | 29.2659 | 0.2503 | varidriv drumspd |
| **4** | 29.8160 | 0.2842 | lower thick varidriv drumspd |
| **3** | 29.9797 | 0.2630 | upper varidriv drumspd |
| **4** | 30.3454 | 0.2790 | upper thick varidriv drumspd |
| **4** | 30.7022 | 0.2755 | lower upper varidriv drumspd |
| **5** | 31.3174 | 0.2891 | lower upper thick varidriv drumspd |
| **2** | 37.9932 | 0.1645 | lower varidriv |
| **3** | 38.8057 | 0.1762 | lower thick varidriv |
| **2** | 39.4653 | 0.1501 | upper varidriv |
| **3** | 39.9300 | 0.1651 | lower upper varidriv |
| **3** | 39.9691 | 0.1648 | upper thick varidriv |
| **4** | 40.6942 | 0.1773 | lower upper thick varidriv |
| **2** | 41.2909 | 0.1321 | thick varidriv |
| **2** | 42.8476 | 0.1168 | lower thick |
| **2** | 43.0135 | 0.1152 | lower drumspd |
| **1** | 43.0523 | 0.0951 | lower |
| **3** | 43.3575 | 0.1315 | lower thick drumspd |

| Number in Model | C(p) | R-Square | Variables in Model |
|---:|---:|---:|:---|
| **1** | 44.0703 | 0.0851 | varidriv |
| **3** | 44.3807 | 0.1214 | lower upper thick |
| **2** | 44.5670 | 0.0999 | upper thick |
| **2** | 44.6876 | 0.0987 | lower upper |
| **3** | 44.8707 | 0.1166 | lower upper drumspd |
| **3** | 44.9002 | 0.1163 | upper thick drumspd |
| **2** | 45.0258 | 0.0954 | upper drumspd |
| **4** | 45.1263 | 0.1337 | lower upper thick drumspd |
| **1** | 45.4837 | 0.0712 | upper |
| **1** | 45.5573 | 0.0705 | thick |
| **2** | 46.5434 | 0.0805 | thick drumspd |
| **1** | 50.9902 | 0.0171 | drumspd |

*It is important to search for* $c_p$ *which is very small and which is close to* $6 + 1 = 7$*.From the diagram above, the model whose* $c_p$ *is close to* $7$ *is ph, upper, varidiv and drumpsd.*

## AUTOMATIC METHODS

*It is important to investigate the model selected with* $c_p$ *procedure, apply* Stepwise selection procedure.

| Summary of Forward Selection | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Step** | **Variable Entered** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | ph | 1 | 0.2913 | 0.2913 | 23.0987 | 20.96 | <.0001 |
| **2** | lower | 2 | 0.1571 | 0.4484 | 9.1183 | 14.24 | 0.0004 |
| **3** | thick | 3 | 0.0214 | 0.4697 | 8.9425 | 1.98 | 0.1661 |

| Summary of Stepwise Selection | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Step** | **Variable Entered** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | ph | | 1 | 0.2913 | 0.2913 | 23.0987 | 20.96 | <.0001 |
| **2** | lower | | 2 | 0.1571 | 0.4484 | 9.1183 | 14.24 | 0.0004 |

*From the diagram above, using forward selection, the model will contain ph, lower and thick. Also, using stepwise selection procedure, it can be said that ph and lower are the onle regressors left in the model.*

## SUMMARY REPORT

*For testing the best model for Solid, the following points should be noted*

- The variable ph has a strong correlation with solid
- Scatter plot of solid with most independent variables are not linear
- With stepwise selection, ph and lower are two variables that remain in the model. Also, with forward selection variable, ph, lower and thick are the only variables that remain in the model.

## *CONCLUSION*

Using all possible procedure, ph is the best single predictor of Solid
Using stepwise regression, the model will contain two variables: ph and lower with the following best of fit

$solid = 4.788 + 2.3078ph - 0.207lower$