**Group Project:**

**Water Usage of Production Plant**

STAT 512

Division 4

Group 8

Tianyi Fang

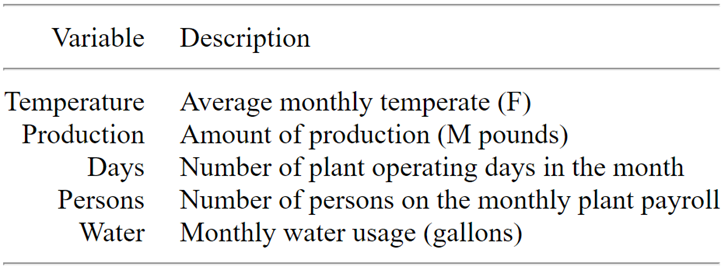
Shangcheng Yang

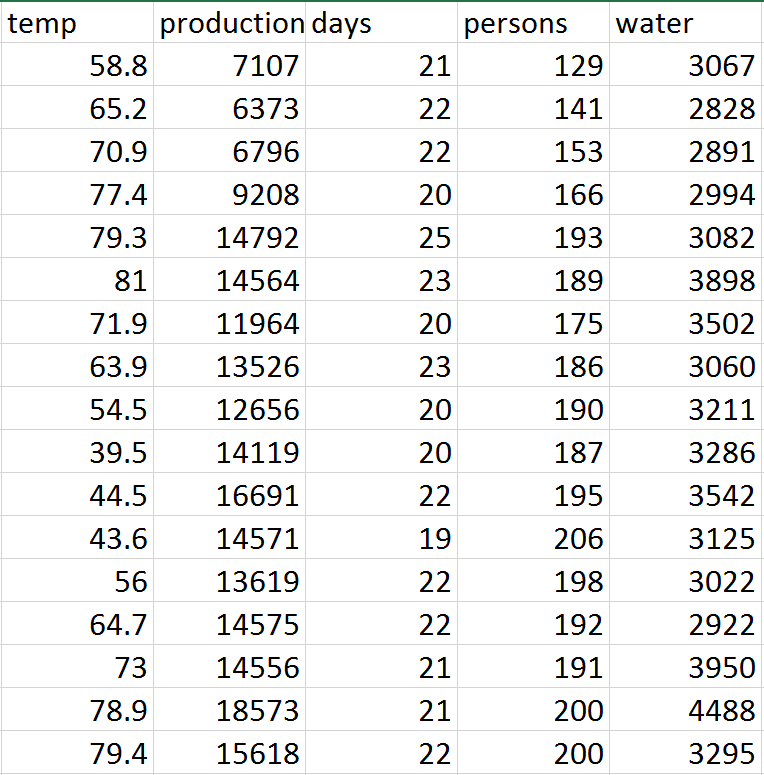
Zhiyuan Jiang

Wonseok Chung

**Introduction**

A production plant cost-control engineer is responsible for cost reduction. One of the costly items in his plant is the amount of water used by the production facilities each month. He decided to investigate **water usage** by collecting seventeen observations on his plant's water usage and other variables.





Our project is aimed to find out the relationship between the amount of water usage and predictors: temperature, operating days, people and production.

**Part 1 Transformation of Variables for the Full Model**

* **Preliminary**
  + **Correlation between variables**

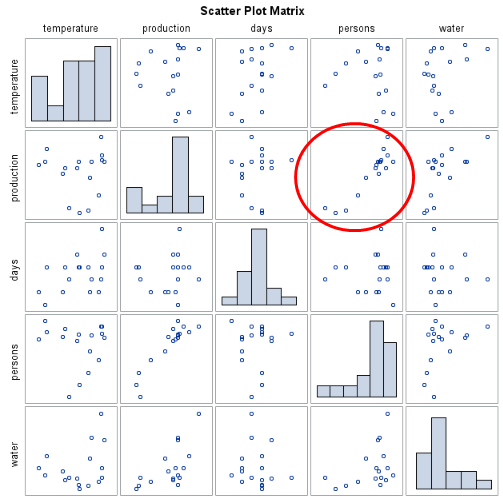


Figure 1. Scatterplot Matrix

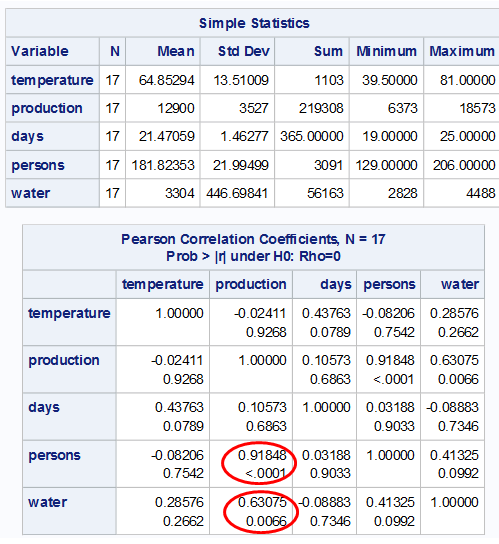


Figure 2. Pearson Correlation Coefficients Matrix

First, looking at the relationships between predictors v.s. predictors and predictors v.s. Response separately by Scatter Plot Matrix (Fig. 1)and Pearson Correlation Coefficients Matrix (Fig 2). For Predictor v.s. Predictor, it is quite obvious that **Production** is highly correlated with **Person**, with correction coefficient 0.91848, which is also supported by the approximately linear relationship of the scatter plot. That is, it is likely that one of them is redundant in predicting Water Usage. We may either introduce an Interaction term of Production and Person to our model or drop one of them. For Predictor v.s. Response, **Production** makes the greatest impact on **Water** Usage with 0.63075, the highest correlation coefficient among all predictors.

**Linearity Check & Transformations**

* **Water vs days**

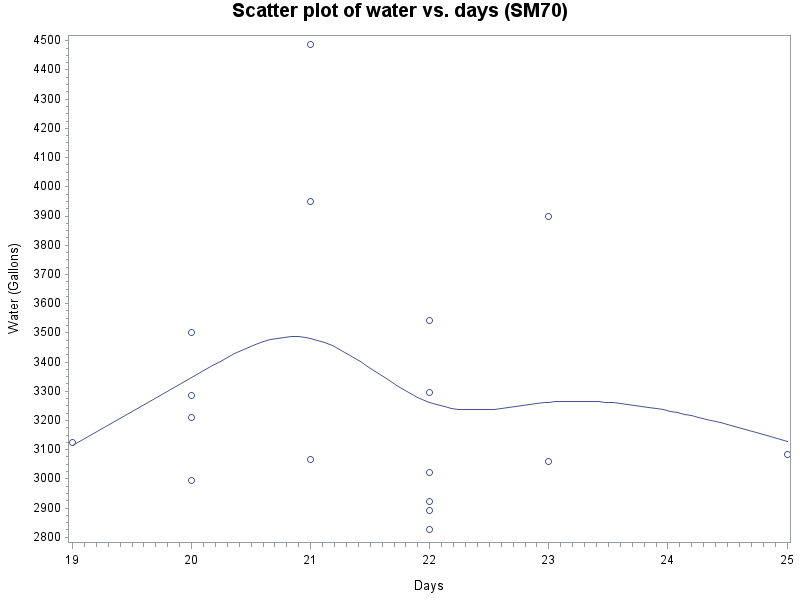


Figure 3 Scatter Plot of Water v.s. Days

The relationship between ‘days’ and ‘water’ is not linear (Fig 3). Either values of days or values of water usages are not normally distributed. For example, there are only one point at days=19 or 25, while 6 points at days=22. Additionally, the ranges of water usage for different days are largely different from each other, which results in curves of the relationship line. However, due to the fact that the predictor ‘day’ is discrete, X transformation is unlikely to make a difference. Thus, Box-Cox procedure is done for water.

Via the Box-Cox analysis (Fig 4), λ in equation of(waterλ = days) is given to be -3. Therefore, using regression model water-3=days to check the linearity between transformed y v.s. predictor, the resulting plot shows as below (Fig 5).

However, problem of linearity between the response variable and predictor seems not to be solved by the Y transformation. According to the correlation matrix attached above, the correlation between ‘water’ and ‘day’ is merely -0.08883. In that case, it would be better off not to make Y transformation because the complexity that follows due to the Y transformation is not worth it regarding to the quite insignificant correlation between **water** and **days**. Thus, the final relationship is

**water = days**

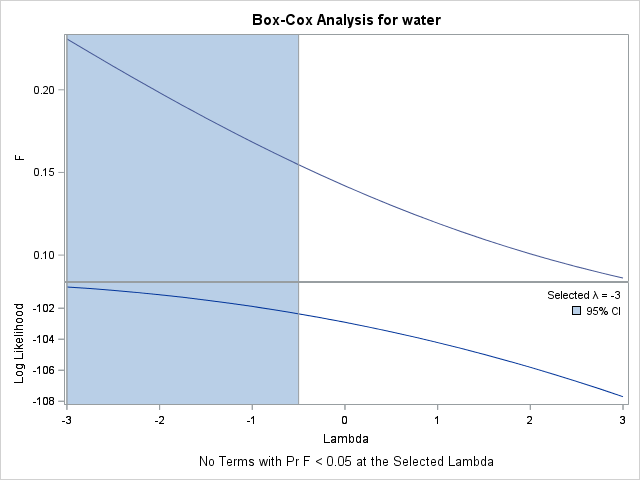


Figure 4 Box-Cox Transformation for water

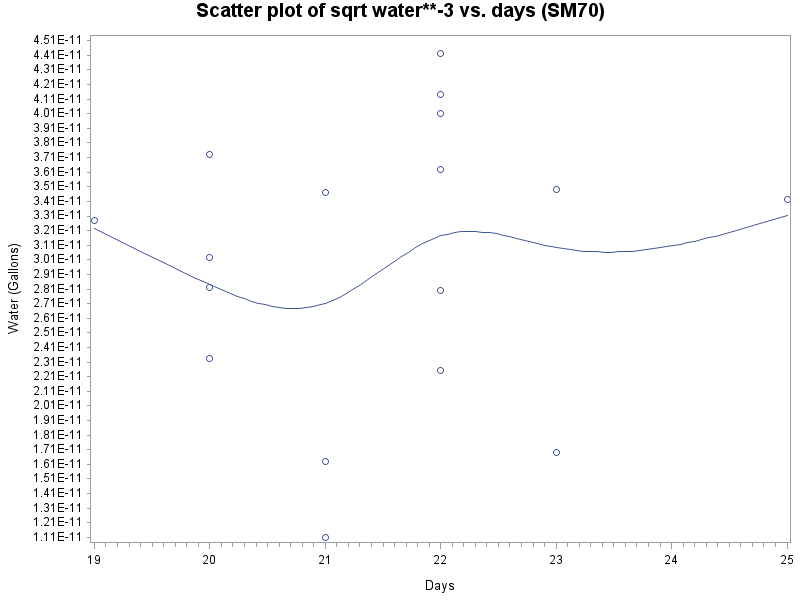


Figure 5 Scatter Plot of water-3 v.s. days

* **Water vs Temperature**

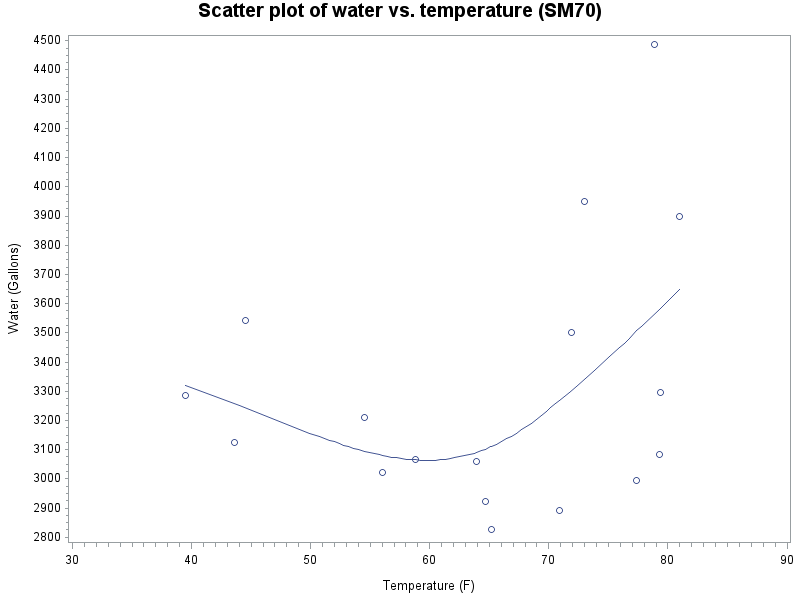


Figure 6 Scatter Plot of water v.s. Temperature

The shape of the relationship between **water** and **temperature** is like that of a quadratic function. (Fig 6) Thus the squared term of ‘temperature’ should be introduced to interpret the plot. Since all values of temperature are positive, there will be a perfect correlation between ‘temperature’ and ‘temperature2.’ In that case, centering method is applied to eliminate the correlation between these two variables. We set 64(mean of temp) as the center of temperature, and subtract it from ‘temperature’ before being squared. Thus, the final relationship is

**water = temperature, (temperature-64)2**

* **Water vs production**

For the relationship between Production and Water Usage, it will much likely play a vital role in the final model and deserves a thorough consideration here, as of the high correlation stated in the beginning of this report. Firstly, a scatter plot is made to illustrate the general relationship, which is shown in the below.

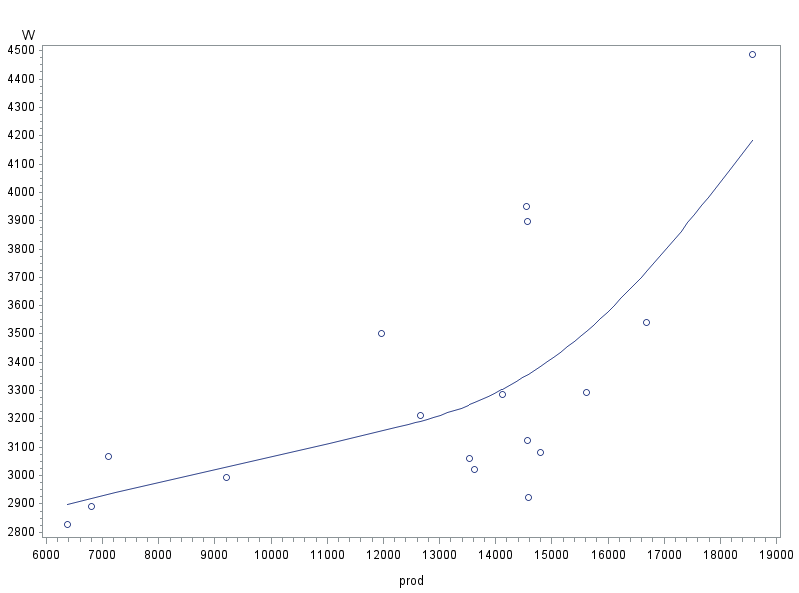


Figure 7. Scatter plot on water versus production

Approximately, the plot (Fig. 7) behaves linear in the first appearance (though it may appear to be loosely quadratic). But to dig into the plot more, it shows that there are two pools of separate data sets, with one pool below ‘9500 prod’, and another above ‘12000 prod’, which seems more reasonable if a piecewise function is used to model the relationship (only up to 2-segment piecewise is recommended for model simplicity). To construct this two-segment piecewise model, a breaking point of ‘12000 prod’ is used in SAS REG and GPLOT function for displaying the piecewise plot, as shown in the Fig. 8. However, it reveals a major drawback of this model that the first segment of the line is poorly supported by the data, with four data points scattering around the left end of the line with a large data-free zone connecting to the second line segment , and the slope of the first part, as a result, is largely affected by the starting point of second half of the line whose shape is more trustworthy, as the two lines need to join each other to form a piecewise model. Correspondingly, the t-test result of the first line segment is very unsatisfactory (p-value 0.76). However the second segment of the piecewise line is acceptable due to the small size of this dataset with only 17 observations, though it may be driving by an influential point on the upper-right corner of the plot and suffers from the problem of a large SSE.

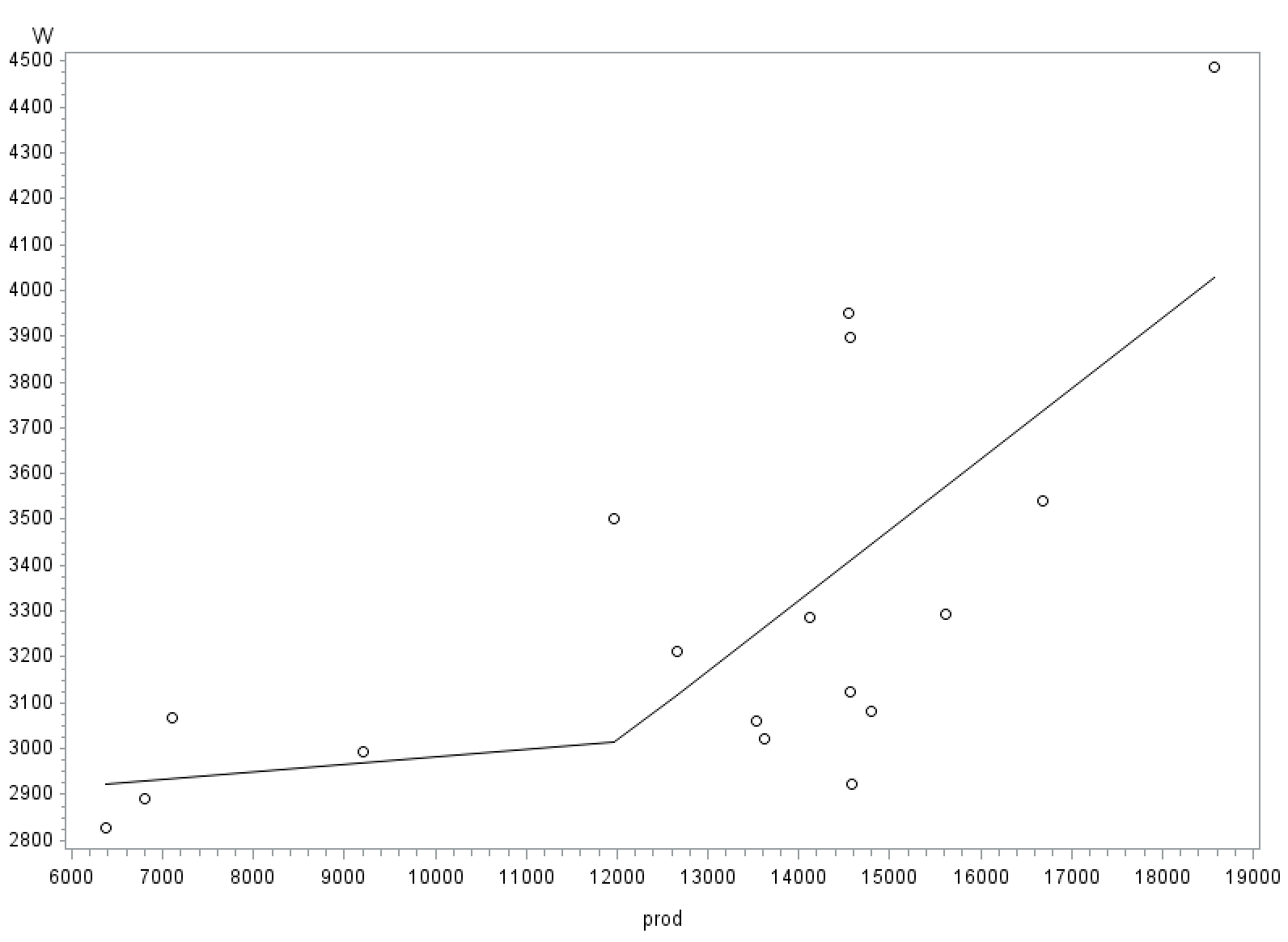
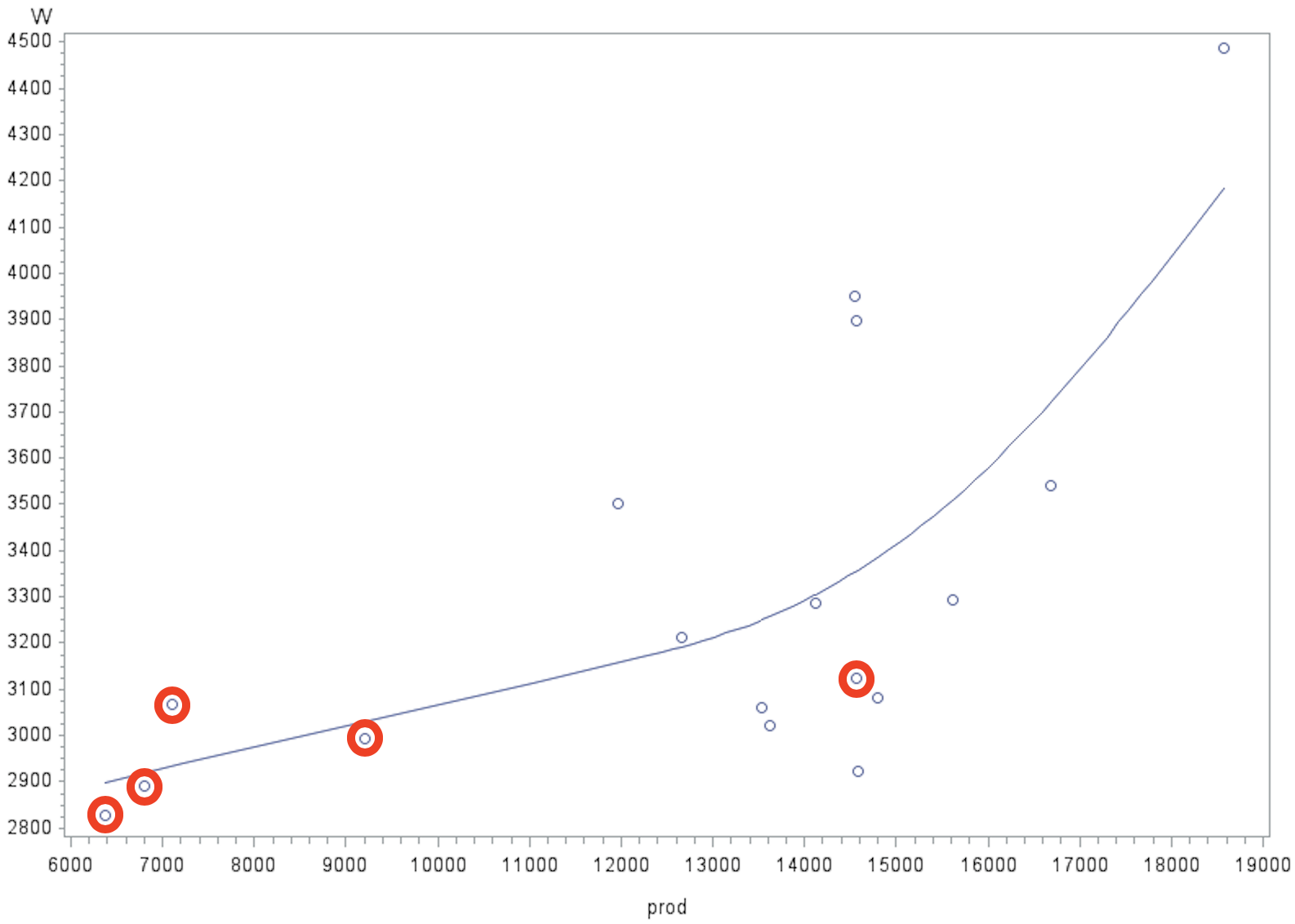
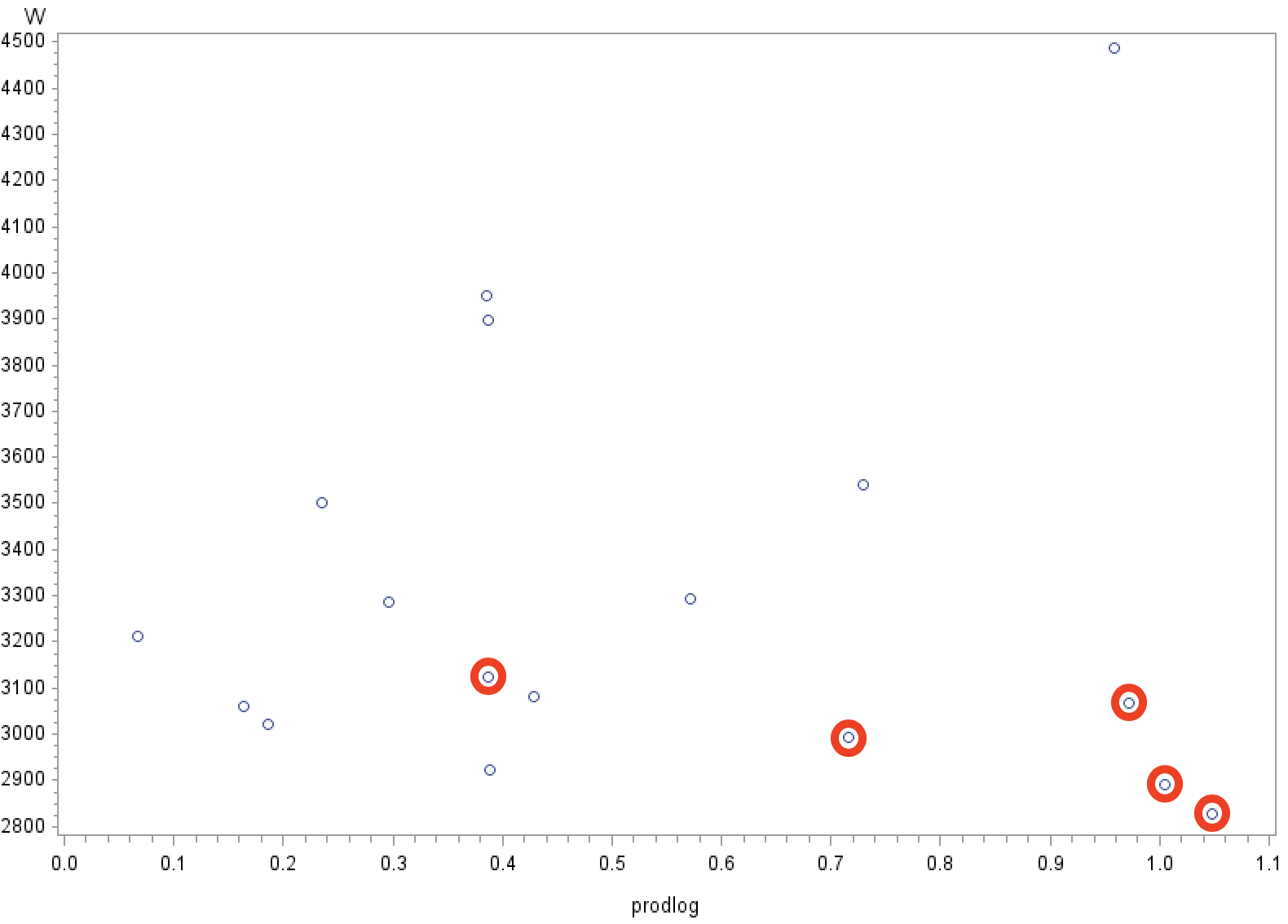


Figure 8. Original piecewise plot on water versus production

Although the piecewise plot (Fig. 8) on explanatory variable - Production shows some problems, it seems plausible to keep this transformation if certain remedies can be taken places to relax them. And to deal with that major problem, it is desirable to shrink the gap ‘9500 prod - 12000 prod’ while at the same time keep the other parts unchanged.

Inspired by the real life experiences, it is a common sense to assume all relationships will work out effectively only within some range, and based on the limited information of the problem background and the correlation matrix(Fig. 2) on all variables, it is proper to predict that the water usage will behave well-correlated with production varying within some production range (i.e. excluding the above four points), or within some equivalent people range, since these two explanatory variables are very high correlated themselves. Not surprisingly, after examining deeper at the four data points that support the first line segment of Fig. 8, it shows that they all have low values on the other explanatory variable - people, and there exists a one-to-one order-preserved mapping for the four data points from production to people, i.e. the four data points contain the lowest four values for both of these two explanatory variables. Another interesting finding is that for the maximum value, 206, on the explanatory variable people, it has a middle-level production, 14571, as well as a similar water usage level with the above four points examined. This means that the five points together (shown by the red dots in the Fig. 9a) can be grouped as one segment of the piecewise line if to use the people as the indicator function for piecewise model on production. The first advantage of using this external order-preserved indicator is that there will be more flexibilities for the production-axis change, to shrink the gap (will be discussed in the next paragraph). And the second advantage is that it enables a data point to be taken away from the second line section (reduces the data complexity) of the piecewise plot to the first section (enhances the reliability).

a b

Figure 9. Scatter plots on specific data points before (a) and after (b) the production std-abs-shift-log transformation

To shrink the gap on production as x variable, the first step is to standardize the production variable through SAS STANDARD with 0 mean and 1 std. Then, the four lowest production points are put aside on the negative axis, while the others (except one) are on the positive axis. Secondly, it is to take the absolute value of all the points, so that to mirror the four points on the positive axis. The third step is to add all points by 1 (shift log(x) to positive values) and take the log function. This is by discovering the fact that the data points cluster in the lower-value axis, and log function can serve as a way to increase the data interval when data values are small and decrease the data interval when data values are big. Figure 9b shows the results after the production transformations and the red dots are the same red dots on Figure 9a. Since the indicator function Ip created is based on the other variable people, it is straightforward to list out the five red data points by people range, marked by the lowest four values and the highest value on people (or explicitly, the people range 170-200 as the normal range, with XIp=1, outside the range Ip=0). Using the external indicator, it is direct to move the normal range 1.1-unit right (adjustable) to create a more appropriate piecewise plot, with a shrunken gap, as the plot below. (Fig. 10)

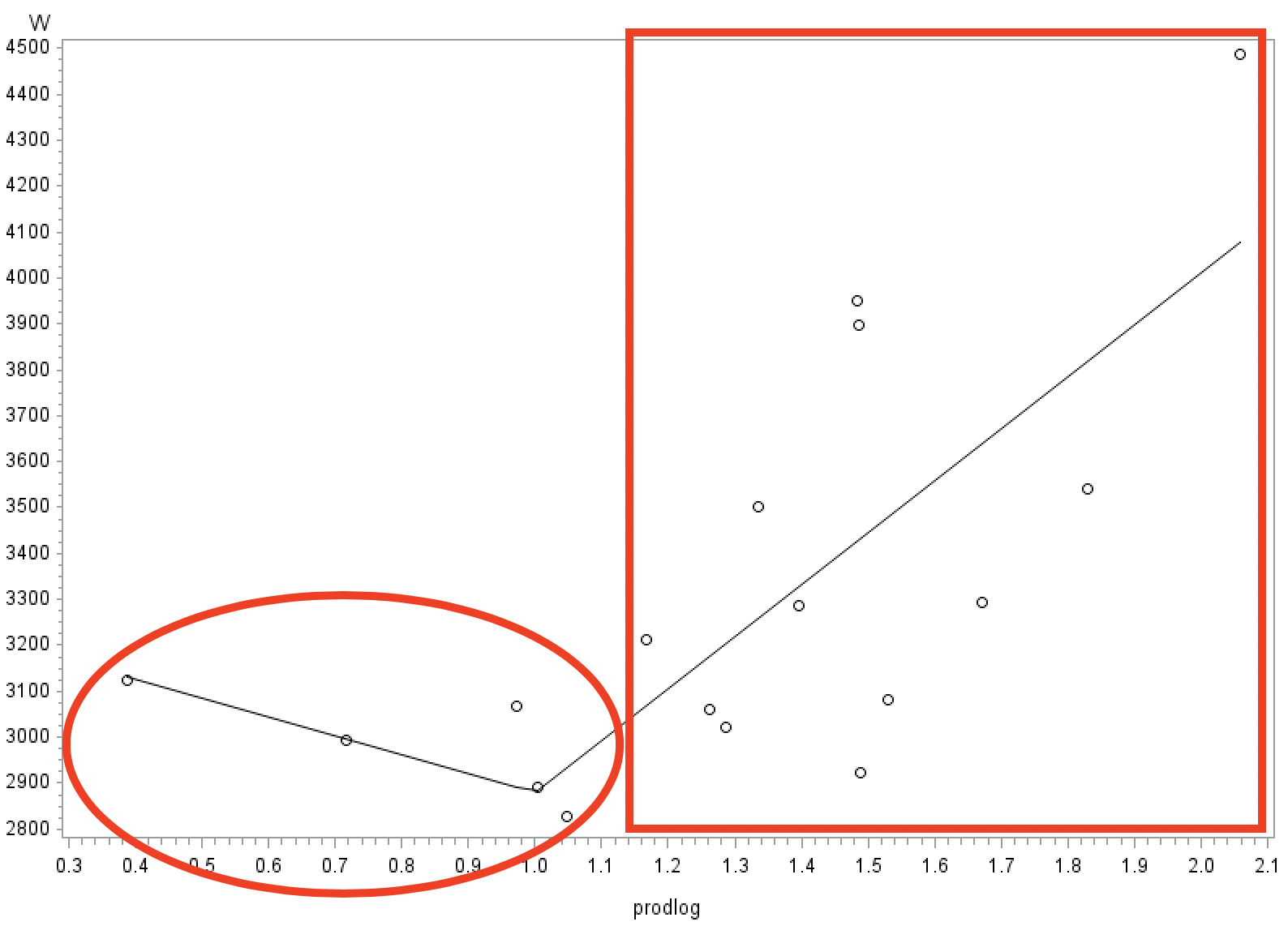
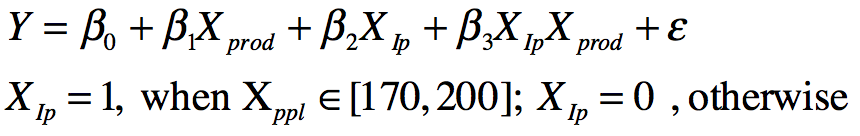


Figure 10. Adjusted production piecewise plot based on people with abnormal range (marked in circle) and normal range (marked in rectangle)

For this production transformation, information is used not only from production itself, but also from people, and the final relationship is formulated as below,



* **Water vs persons**

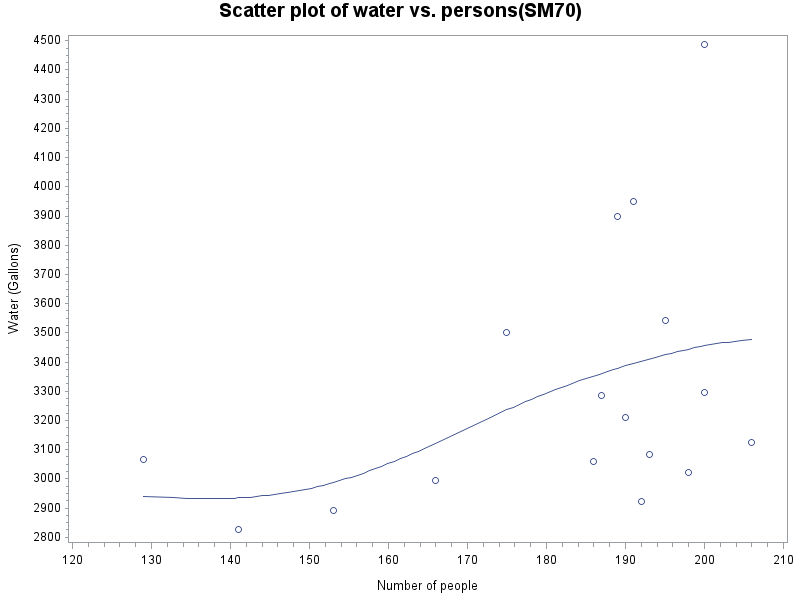


Figure 11 Scatter Plot of water v.s. person

The relationship between ‘water usage’ and ‘persons’ seems approximately linear. However, as the correlation of ‘production’ and ‘persons’ is extremely high, there is no significance transforming ‘persons’ that is mostly explained already by ‘production.’ Thus, it is appropriate to decide not to transform the predictor ‘persons’ and leave it as it is to avoid any complexity. The final relationship is

**water = person**

Therefore, finalized model after all the transformations is,

**water usage = β0 + β1\*days + β2\*temperature + β3\*(temperature-70)2 + β4\*Ip + β5\*log(abs(production)+1) + β 6\*(log(abs(production)+1)\*Ip) + β7\*persons**

**Part II Selection of best model**

* **Cp Criterion**

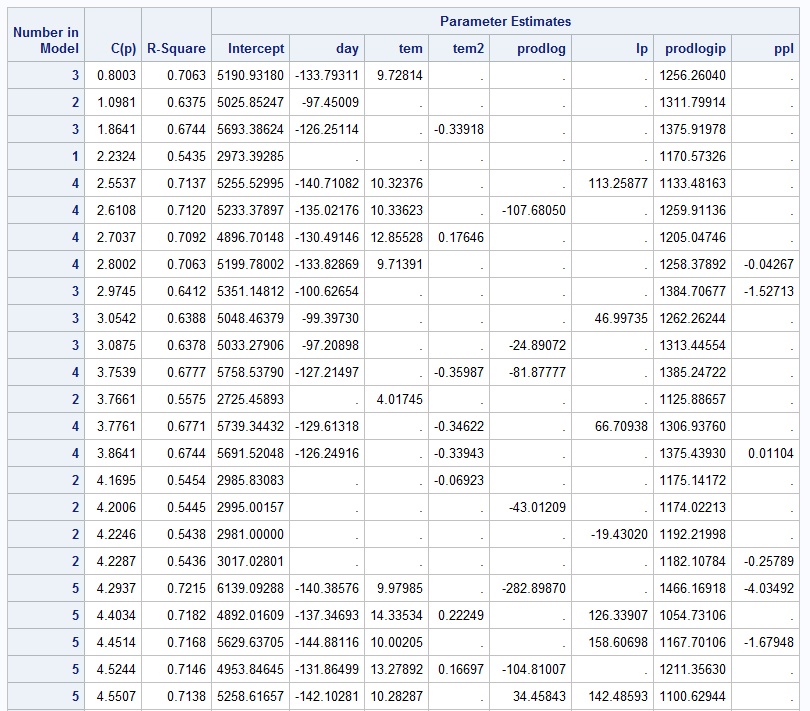


Figure 12 Cp Criterion Table

According to Cp Criterion: Cp≤p and minimizing Cp, we adopt **the first subset model** in the table after deliberately consideration and comparison between the last subset model in the table, which includes all the piecewise variables of predictor Production, and the first subset model in the table with simply ‘day’, ‘tem’, and ‘prodlogip’. Although the first subset model, calling it model A, dropped the piecewise components ‘prodlog’ and ‘Ip,’ the C(p) is significantly lower than any other subset models and there is not much difference with the R2 values. Thus, it is better off to choose model A, **‘water usage’ = ‘day’ + ‘tem’ + ‘prodlogip’**

* **Stepwise method**

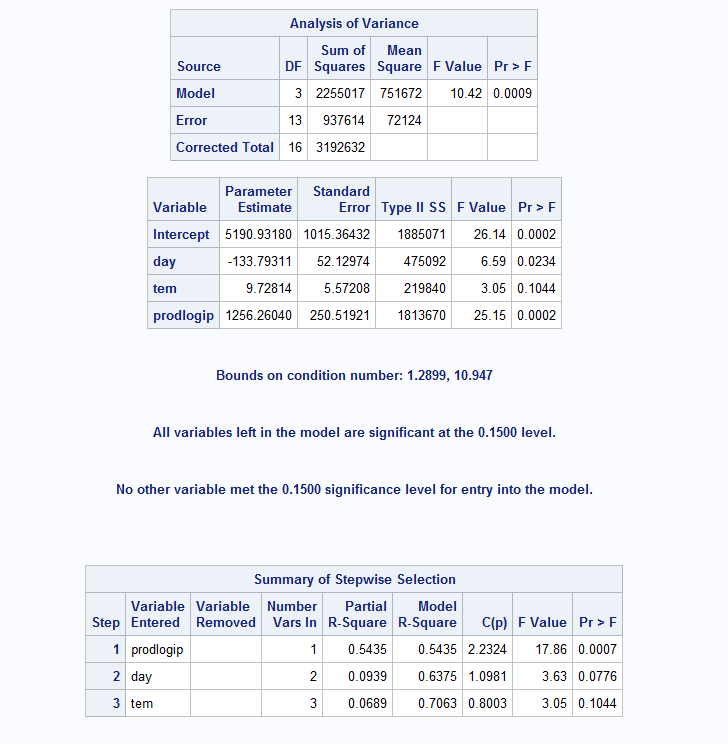


Figure 13 Steps and results of Stepwise method

Now here is the subset model chosen by the stepwise method. At first, ‘prodlogip’ entered the base null model with F-value = 17.86 and p-value = .0007. Predictor ‘day’ followed to enter the model with F-value = 3.63 and p-value = .0776. After so, none of the two predictors present in the model was taken out by the backward elimination. Lastly, predictor ‘tem’ entered the model and again, no elimination of the three variables present in the subset model was excluded. Therefore, the finalized subset model via stepwise method is **‘water usage’ = ‘day’+ ‘tem’+ ‘prodlogip,’** identical to the subset model obtained by the C(p) criterion.

**Part III Diagnostics**

**Assumptions of model**

Best models from question 2 and 3 are identical, **Water Usage = day tem prodlogip.**

Check the assumptions of model.

* **Linearity:**

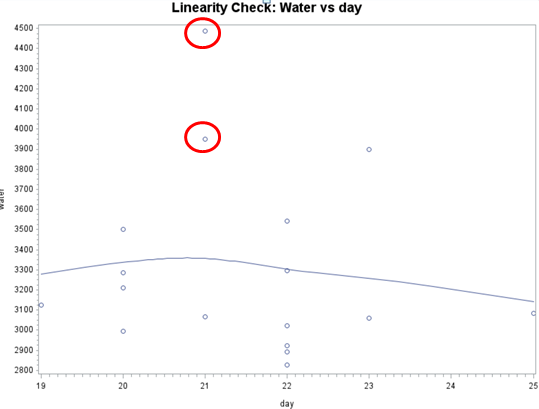


Figure 14. Linearity Check: Water vs. Day

First, plot Water vs. Day (Fig. 14) to check the linearity relationship. From the graph, notice that it is approximately linear, however there is a inflexion when day at 21. It is due to the reason that when day is equal to 21, there are two points which the water usage is much higher than the average, so line is pulled up. Meanwhile, when day is equal to 22, there are 4 points which is below the average water usage, so the line is pulled down.

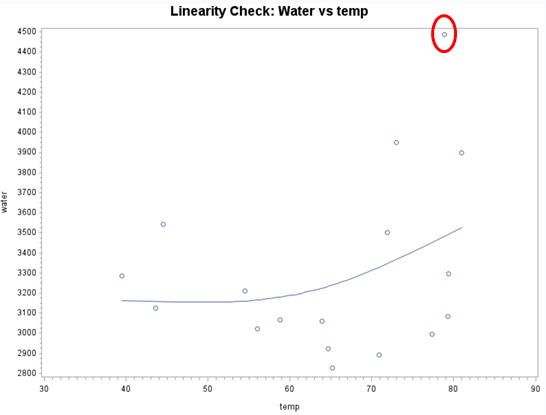


Figure 15. Linearity Check: Water vs. temp

Secondly, use Water vs. Temp(Fig. 15) plot to check the linearity. The plot is approximately linear, while there is an outlier at temperature 79, which pulls the line up.

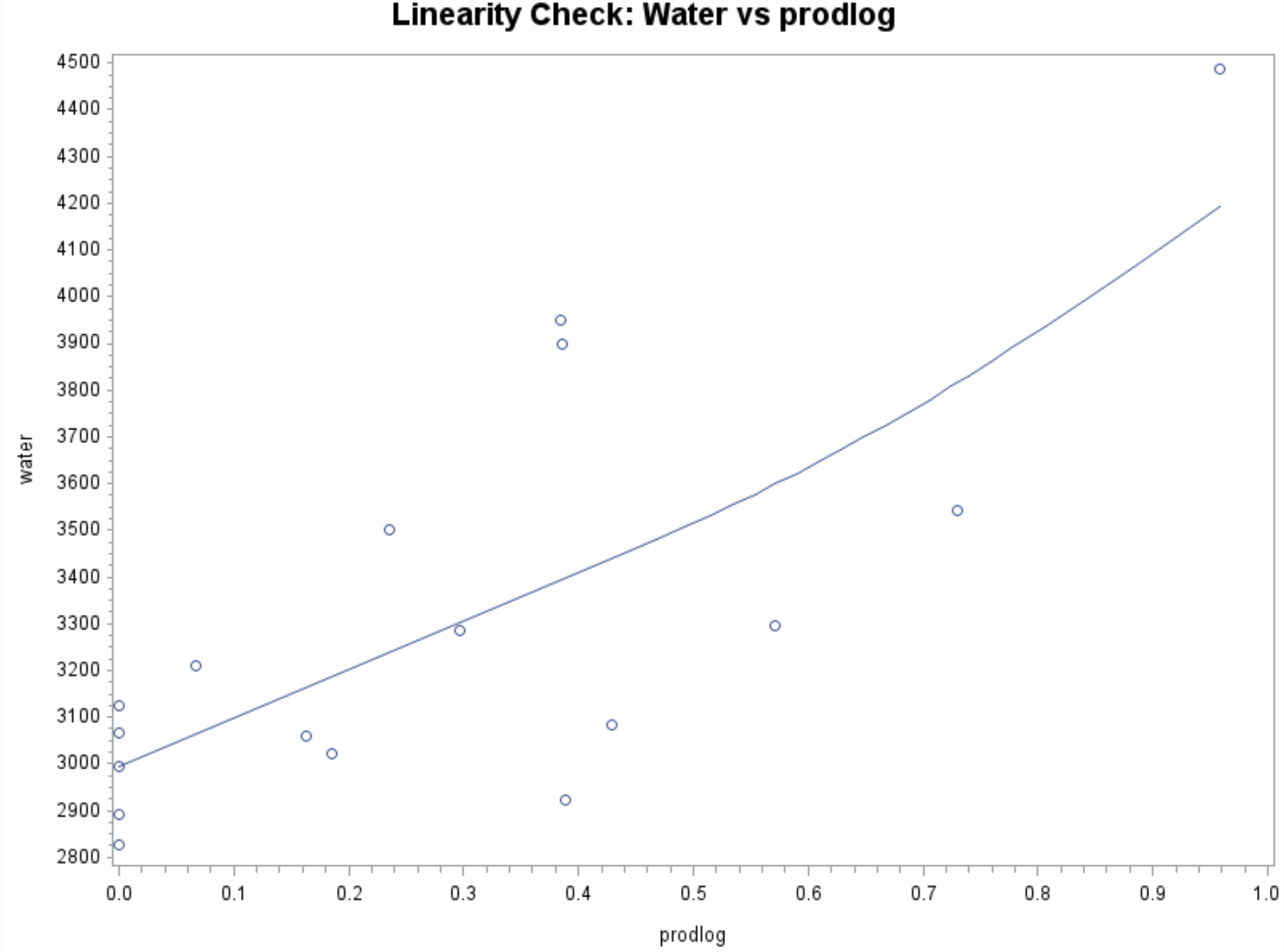


Figure 16. Water vs. prodlog

Thirdly, Check the linearity between water and prodlog. This plot (Fig. 16) is perfectly linear, thus assure that the relationship between water and prodlog is linear, which is reasonable in real world experience, since the more productions are produced, the more water is needed.

* **Constant variance:**

Use plots of residuals vs. X’s (explanatory variable) to check the assumption of constant variance.

In the plot of residuals vs. day, the residuals roughly form a ‘horizontal band’ around 0, thus it indicated that the variance is constant. However, the plot of residuals vs. prodlog show that the constant variance assumption is violated, since when prodlog at 0.4, the range of residuals is much larger. Besides, the residuals vs. temp plot seems to be a megaphone shape, thus it also violated the assumption of constant variance.

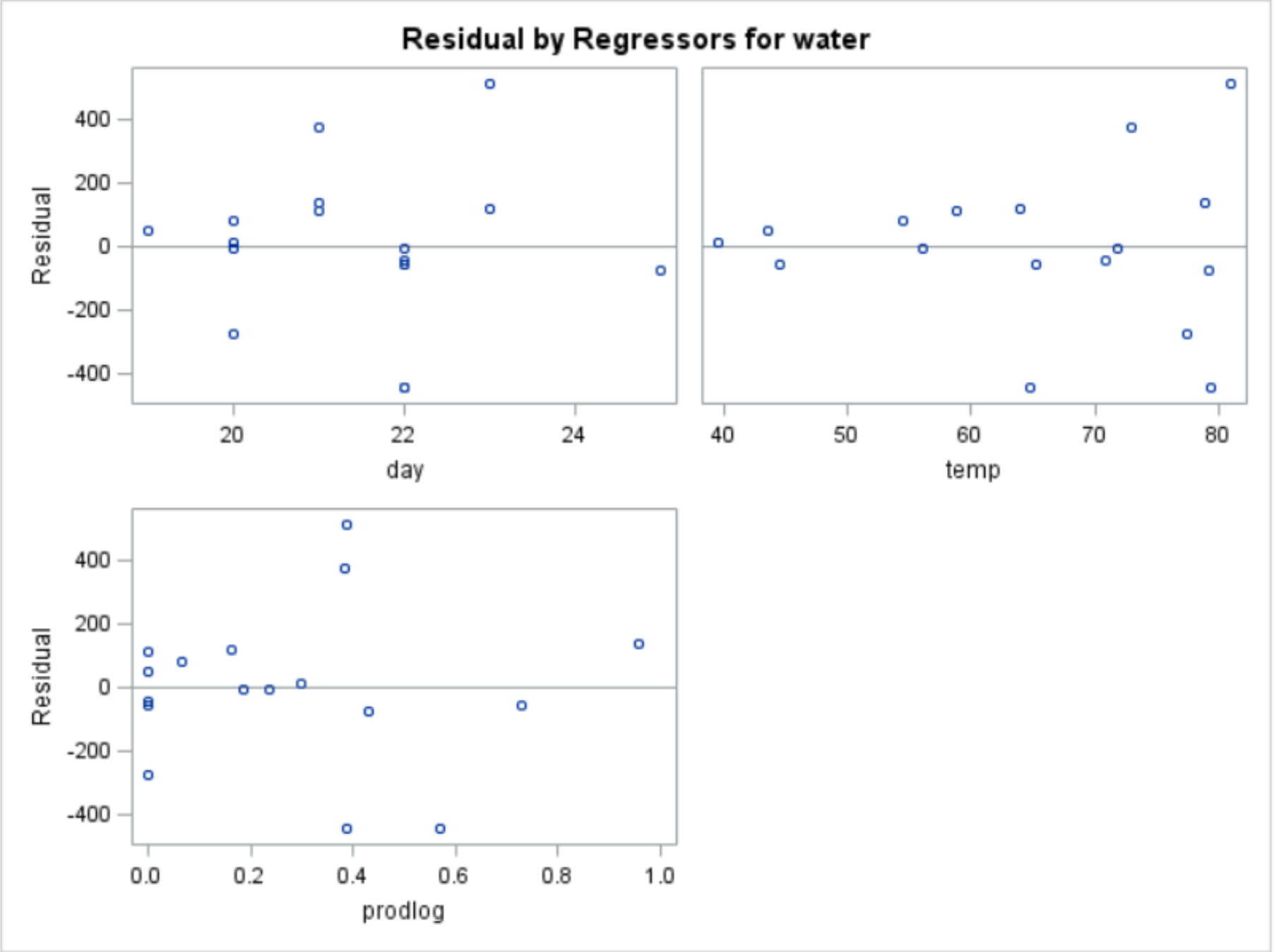


Figure 17. Residuals vs. X’s

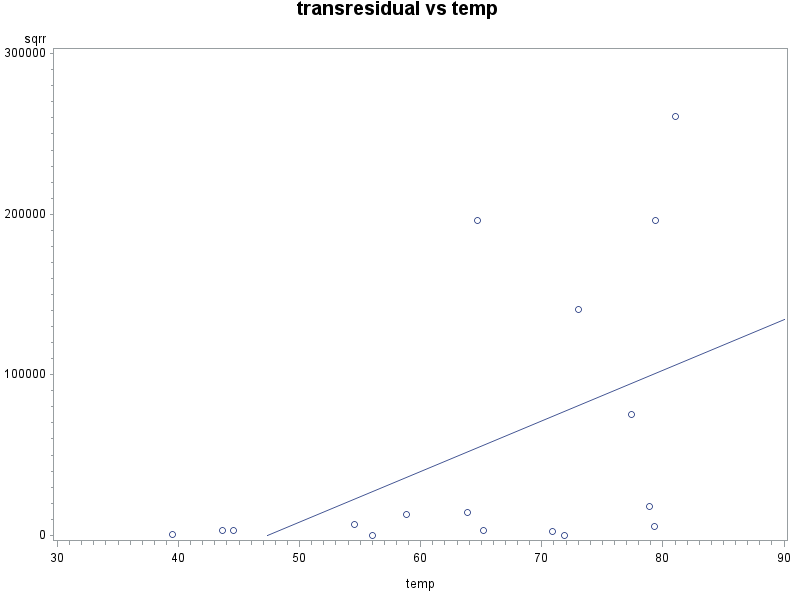
* **Remedies Measures**

Weighted Regression

i) For temperature:

Plot the absolute residuals vs. *temp* and the squared residuals vs. *temp*.

From the graphs below, the absolute value of the residuals (Fig. 18b) appears to have a fairly linear relationship with *temp* (it appears more linear than does the graph of squared residuals vs. *temp* (Fig. 18a)). Thus, we will model standard deviation as a linear function of *temp*. Model the absolute residuals as a function of *temp,* and use the predicted values of that regression as weights.

Figure 18a. Squared residuals vs. temp

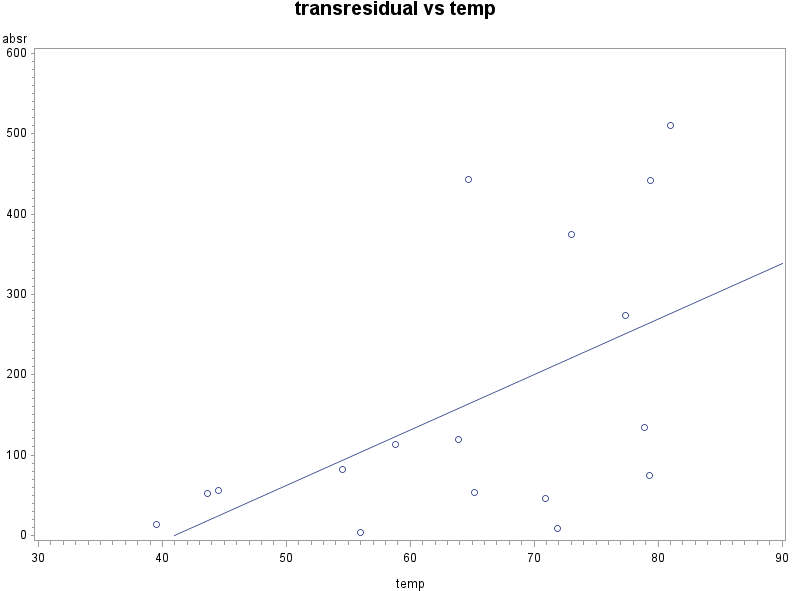


Figure 18b. Absolute residuals vs. temp

After that although the variance is still non-constant (Fig. 19), it is much better than the origin one.

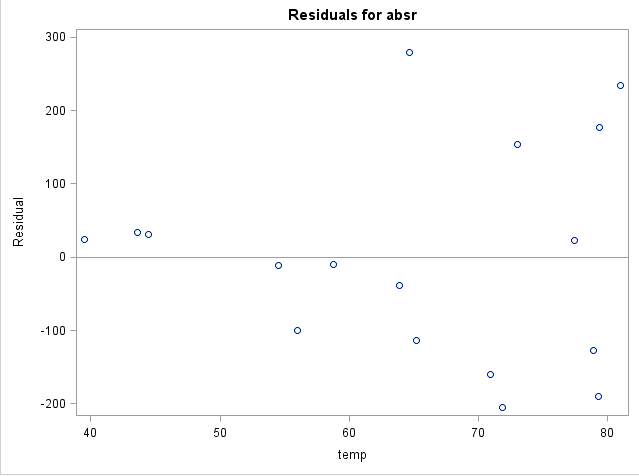


Figure 19. Weighted Regression residuals vs. temp

ii) For prodlog:

Plot the absolute residuals vs. prodlog and the squared residuals vs. prodlog.

From the graphs below, the absolute value of the residuals (Fig. 20a) appears to have a fairly linear relationship with prod (it appears more linear than does the graph of squared residuals vs. prod (Fig. 20b)). Thus we will model standard deviation as a linear function of prodlog. Model the absolute residuals as a function of prodlog, and use the predicted values of that regression as weights.

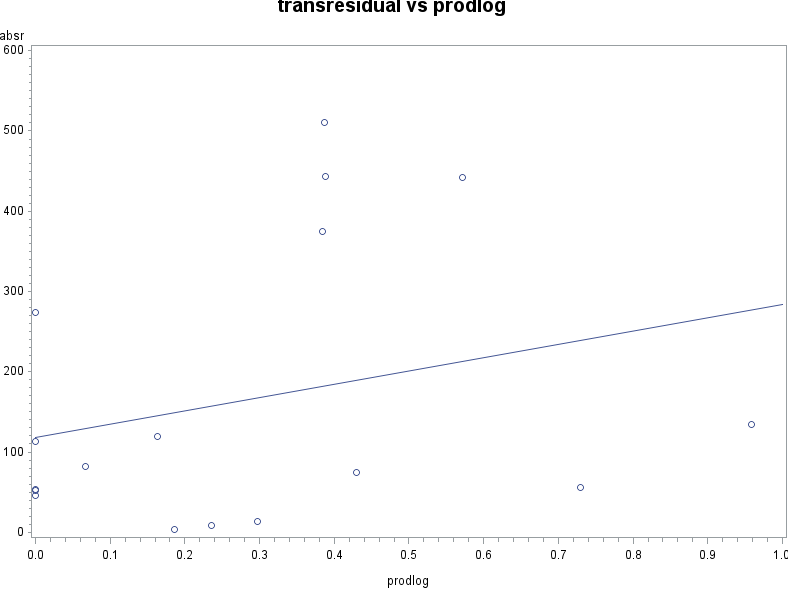


Figure 20a. Absolute residuals vs. temp

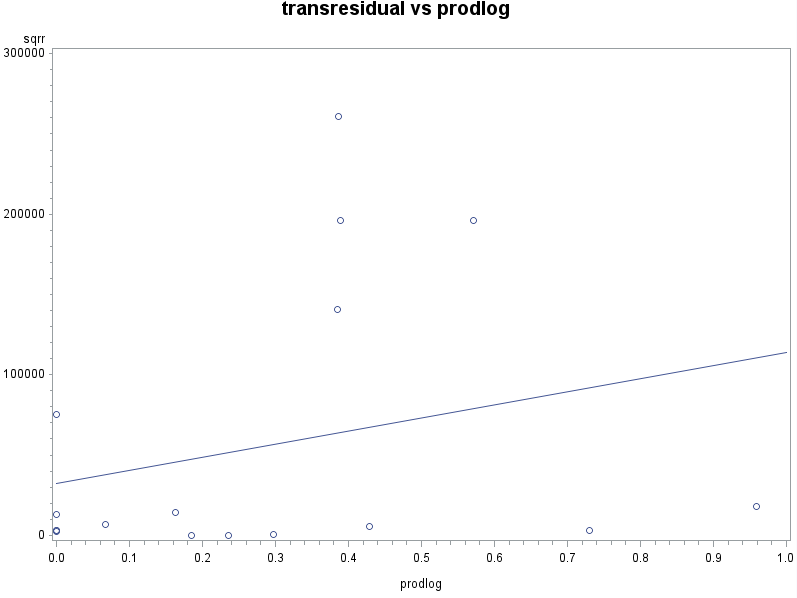


Figure 20b. Squared residuals vs. temp

After that the variance (Fig. 21) is still remain non-constant.

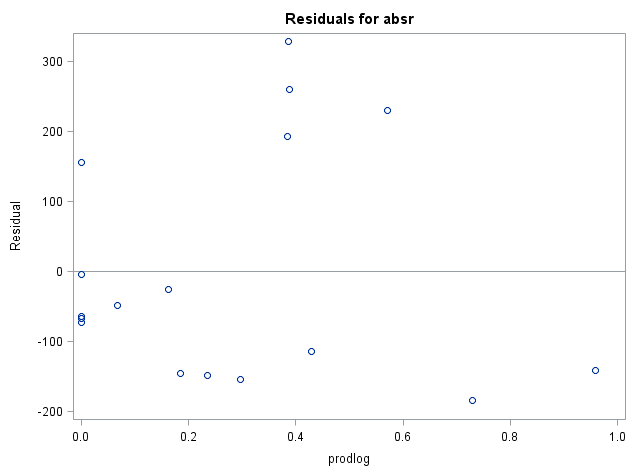


Figure 21. Weighted Regression residuals vs. prodlogip

* **Normality:**

Use QQ-plot and histogram to check the normality(Fig. 22). Both graphs shows that the residuals approximately follow the normal assumption.

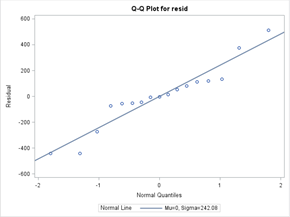
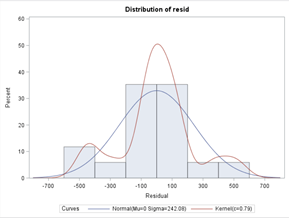
 

Figure 22. QQ plot(left) and Histrogram(right)

* **Independence:**

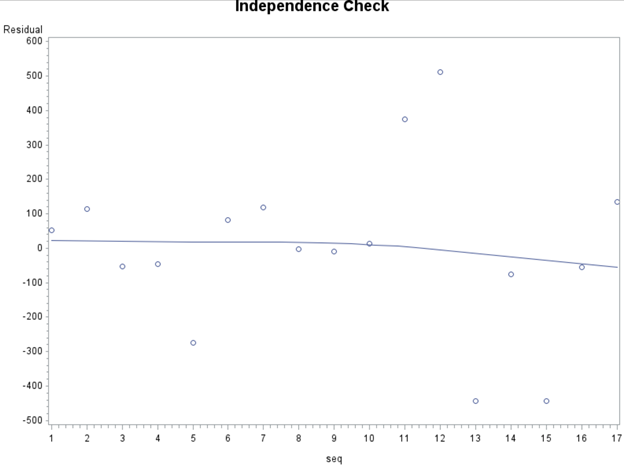
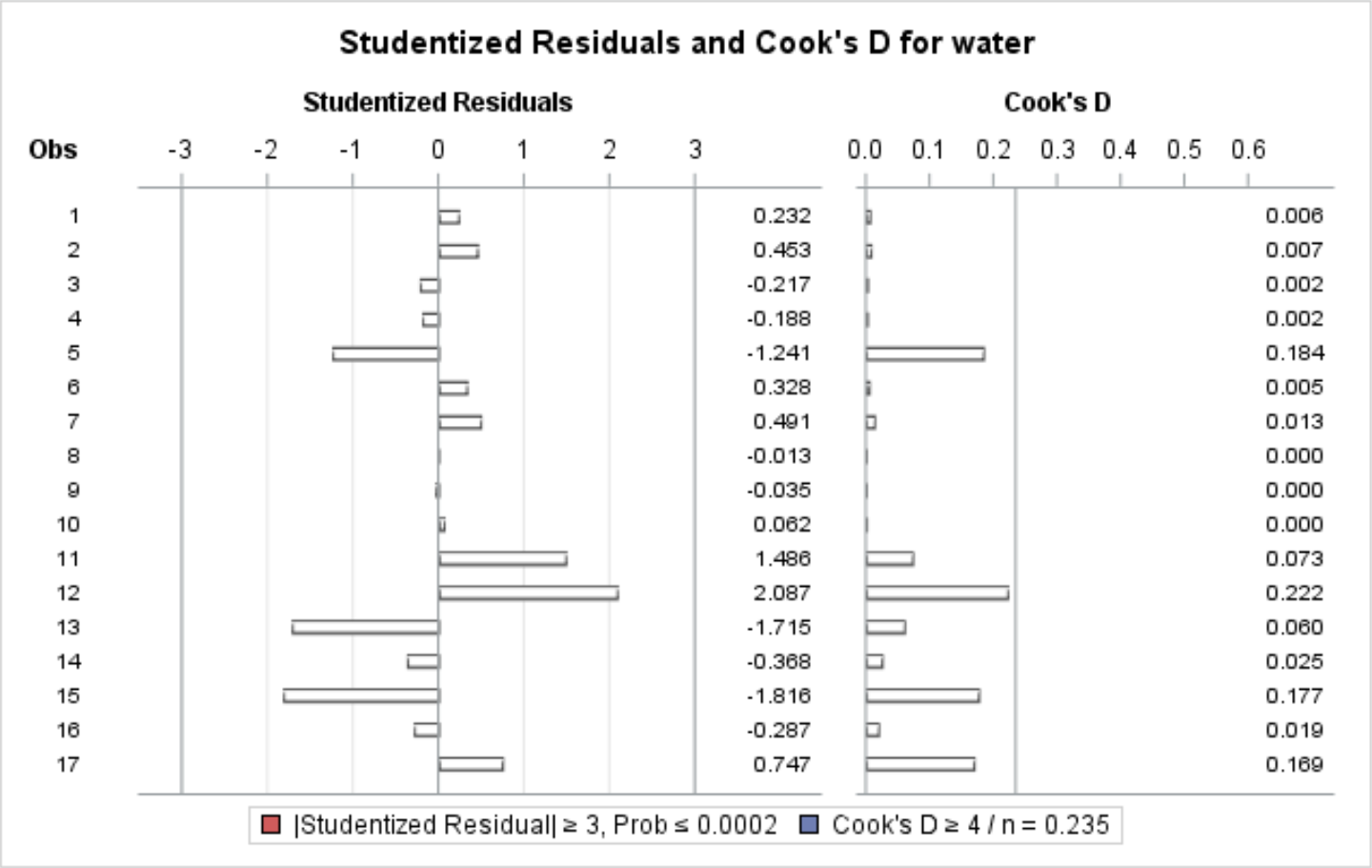


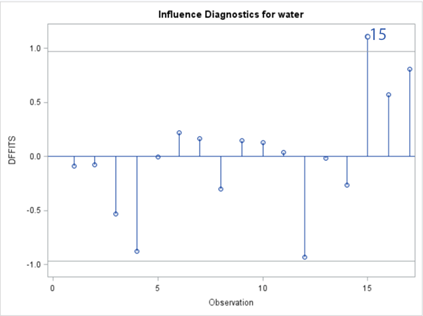
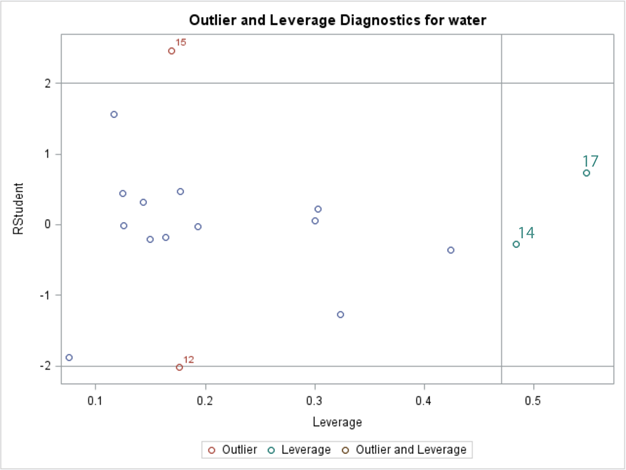
Figure 23. residuals vs. sequence

Check independence assumption by using plot of sequence vs time(Fig. 23). By observing the graph, the sequence plot of residuals seems to behavior in a random manner, and there is no obvious pattern. Thus, it follows independence rule.

**Influence of data points**



(a)

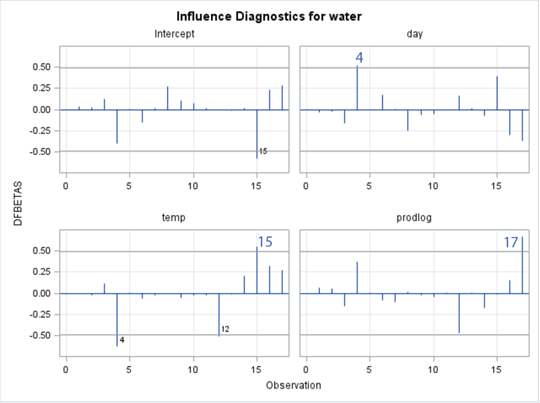


(b) (c)

Figure 24 Influence Check of predictors:

(a)Studentized Residuals & Cook’s D (b) Studentized Deleted Residuals

(c) DFFITS (d) DFBETAS



(d)

* **Studentized Residuals:**

There is **no observation** has a Studentized Residual larger than 3, so no observations need to be checked according to Student Residuals.

* **Studentized Deleted Residuals:**

Testing whether the case with the largest studentized residual is an outlier. t(n-p-1,1-α/2n)=From the SAS result, **Obs#15** and **#12** seem to be outliers.

* **Cook’s D:**

Measuring the influence of case i on all of the Ŷi’s and large values suggest an observation of significant influence. Since there is no observation has Cook’s D larger than F5,17-5 (0.5)=0.92 of our model, there is **no observation** has significant influence on Ŷ, need to be checked according to Cook’s D.

* **Hat matrix diagonals:**

Measuring how much Yi is contributing to the prediction of Ŷi. Observations with extreme values for the predictors will have more influence. 2p/n of our model is 2\*4/17=0.4705. Since there is **no observation** has Hat Diag H larger than 0.588, no obs need to be checked.

* **DFFITS:**

Another measure of the influence of case i on its own fitted valued Ŷi. Value larger than

2\*sqr(p/n), which in our model is 0.9701. Since this is a small dataset we would use 1). **Obs #15** has DFFITS value of 1.11, which may have some influence on the prediction of Ŷi.

* **DFBETAS:**

Measuring the influence of case i on each of the regression coefficients. Values larger than 2/sqr(17)=0.485 are considered influential. **Obs #4, #11, #17** appears to be influential.

According to the above results of measuring influential observations, Observation #15 shows relatively high frequency in influencing data check, and is the first candidate to be dropped from our final data set due to its significant degree of influence. However, since there is only 17 observations in our model, we’d better keep everyone in our model and treat every data point as meaningful.

* **Measurement of multicollinearity:**

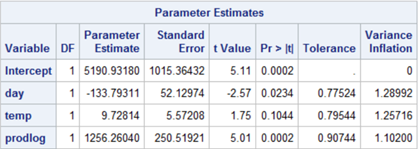


Figure 25 Output of multicollinearity measurement

Tolerance<0.1, the same as VIF>10 means excessive multicollinearity. From the result, no variable’s tolerance is smaller than 0.1, indicating that none of these variables can be predicted by other variables. Thus, there is **no obvious multicollinearity** in our model.

* **Partial Residual Plot**

This plot helps us figure out the net effect of Xi and Y, given other variables are in the model. Since the slope indicating its regression coefficient, a linear pattern in Partial Residual Plot shows that the variable would be useful in the model. All the slopes of our variables seems to be not flat, which means that all variables should be included as a linear term in our model.

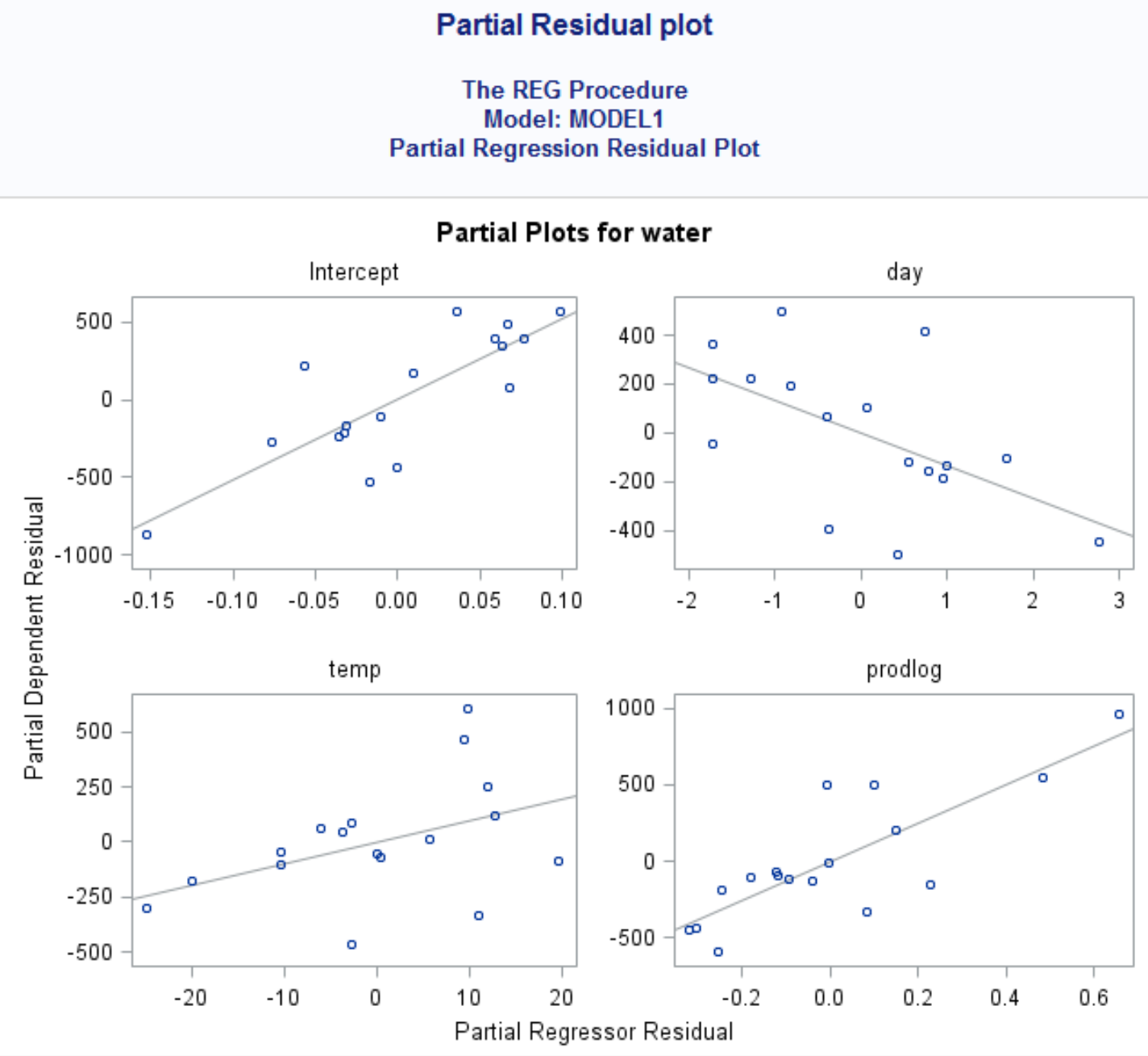


Figure 26 Plots of partial residuals plot

**Part IV Prediction and Summary**

* Model: **Water=5190.93180-133.79311\*day+9.72814\*temp+1256.2604\*prodlog\*Ip**
* 90% Confidence Interval for the mean of the response variable

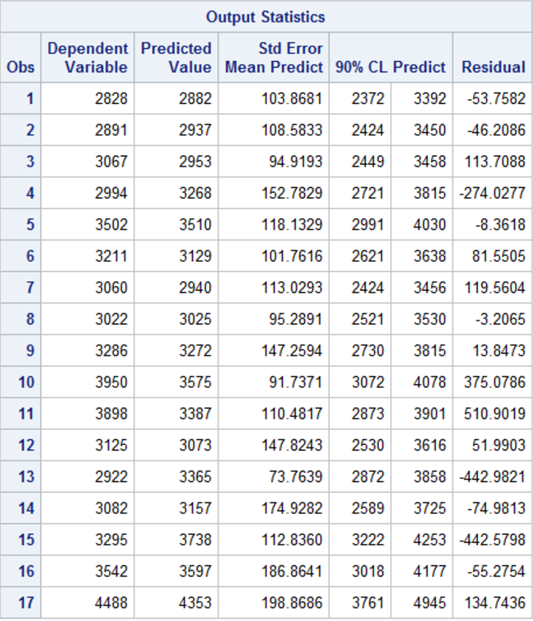
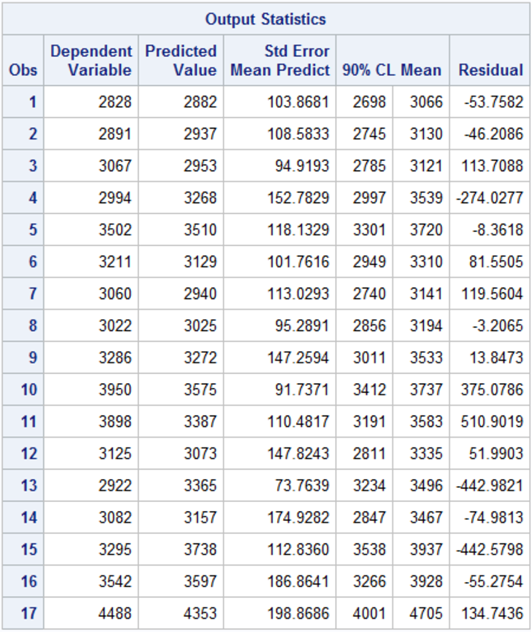


Figure 27(a) 90% Confidence Interval for the mean of response (b) 90% Prediction Interval for the mean of response

* 90% Prediction Interval for individual observations
* 90% Confidence Interval for the regression coefficients

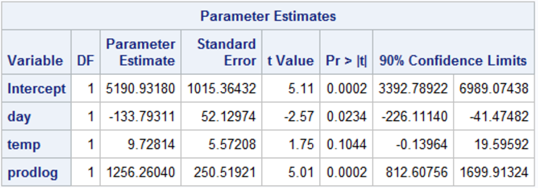


Figure 28 90% Confidence Interval for regression coefficients

**CODE:**

**Question1 Full model**

**data** test;

input temp prod days persons water;

cards;

58.8 7107 21 129 3067

65.2 6373 22 141 2828

70.9 6796 22 153 2891

77.4 9208 20 166 2994

79.3 14792 25 193 3082

81.0 14564 23 189 3898

71.9 11964 20 175 3502

63.9 13526 23 186 3060

54.5 12656 20 190 3211

39.5 14119 20 187 3286

44.5 16691 22 195 3542

43.6 14571 19 206 3125

56.0 13619 22 198 3022

64.7 14575 22 192 2922

73.0 14556 21 191 3950

78.9 18573 21 200 4488

79.4 15618 22 200 3295

;

**proc** **print** data=test;

**run**;

\*get basic information about xi,the relationship of xi and y;

**proc** **univariate** data=test;

var temp prod days persons water;

histogram temp prod days persons water/normal kernel(L=**2**);

**run**;

symbol1 v=cicle i=sm70;

\*linear test of water vs temp;

title"temp vs water";

axis1 label=(angle=**90**'water');

axis2 label=('days');

**proc** **sort** data=test;by temp;

**proc** **gplot** data=test;

plot water\*temp/vaxis=axis1 haxis=axis2;

**run**;

\*linear test of water vs prod;

title"prod vs water";

axis1 label=(angle=**90**'water');

axis2 label=('prod');

**proc** **sort** data=test; by prod;

**proc** **gplot** data=test;

plot water\*prod/vaxis=axis1 haxis=axis2;

**run**;

\*linear test of water vs days;

title"days vs water";

axis1 label=(angle=**90**'water');

axis2 label=('days');

**proc** **sort** data=test; by days;

**proc** **gplot** data=test;

plot water\*days/vaxis=axis1 haxis=axis2;

**run**;

\*linear test of water vs persons;

title"persons vs water";

axis1 label=(angle=**90**'water');

axis2 label=('persons');

**proc** **sort** data=test; by persons;

**proc** **gplot** data=test;

plot water\*persons;

**run**;

\*get correlation of xi and y;

title"correlation of xi and y";

**proc** **corr** data=test noprob plot=matrix(histogram);

var temp prod days persons water;

**run**;

\*water has large relationship with prod&person,

prod has large relationship with person;

\*transformation of days,using ;

**proc** **tranreg** data=test;

model boxcox(days)=identity(water);

**run**;

\*test trans;

**data** transday;

set test;

waterd=water\*\*-**3**;

**run**;

title"Scatter plot of water^-3 vs days";

axis1 label=(angle=**90**'water');

axis2 label=('days');

**proc** **sort** data=transday;by days;

**proc** **gplot** data=transday;

plot waterd\*days/vaxis=axis1 haxis=axis2;

**run**;

\*trans temp,quadratic = temp+(temp-tba)^2;

**data** transtemp;

set test;

temp2=(temp-**70**)\*\***2**;

**run**;

title"Scatter plot of water vs temp^2";

axis1 label=(angle=**90**'water');

axis2 label=('temp');

**proc** **sort** data=transtemp;by temp;

**proc** **gplot** data=transtemp;

plot water\*temp2/vaxis=axis1 haxis=axis2;

**run**;

\*trans prod, need to introduce Ip indicator;

\*days=days, persons=persons,temp=[temp, (temp-utemp)^2],prod=Ip\*trans;

**data** project;

input It Id Ip person temp day prod water;

cards;

0 1 0 141 65.2 22 6373 2828

1 1 0 153 70.9 22 6796 2891

0 1 1 192 64.7 22 14575 2922

1 0 0 166 77.4 20 9208 2994

0 1 1 198 56 22 13619 3022

0 1 1 186 63.9 23 13526 3060

0 1 0 129 58.8 21 7107 3067

1 1 1 193 79.3 25 14792 3082

0 0 0 206 43.6 19 14571 3125

0 1 1 190 54.5 20 12656 3211

0 1 1 187 39.5 20 14119 3286

1 1 1 200 79.4 22 15618 3295

1 1 1 175 71.9 20 11964 3502

0 1 1 195 44.5 22 16691 3542

1 1 1 189 81 23 14564 3898

1 1 1 191 73 21 14556 3950

1 1 1 200 78.9 21 18573 4488

;

\*standardize prod to N(0,1);

**proc** **standard** data=project out=stdprod mean=**0** std=**1**;

var prod;

**run**;

title"standardize prod to std normal";

**proc** **print** data=stdprod;

**run**;

\*trans stdprod to exp/log;

**data** transprod;

set stdprod;

\*prodexp=exp(prod)\*Ip;

prodlog=log(abs(prod)+**1**)\*Ip;

prodlog1=log(abs(prod)+**1**);

**run**;

title"Variable selection via Cp criterion of prod";

**proc** **reg** data=transprod;

model water= prodlog prodlog1 Ip/

selection= cp b;

**run**;

**proc** **print** data=transprod;

**run**;

\*plot stdprod vs water;

title"plot transprod vs water";

axis1 label=(angle=**90**'water');

axis2 label=('prodexp');

axis3 label=('prodlog');

**proc** **sort** data=transprod; by prod;

**proc** **gplot** data=transprod;

plot water\*prodexp/vaxis=axis1 haxis=axis2;

plot water\*prodlog/vaxis=axis1 haxis=axis3;

**run**;

\*regression of all transformed Xi;

**data** projectnew;

set transprod;

temp2=(temp-**70**)\*\***2**;

**run**;

**proc** **print** data=projectnew;

**Run**;

**Question2 Select the best model via Cp Criterion**

title"Variable selection via Cp criterion";

**proc** **reg** data=projectnew;

model water=day temp temp2 prodlog prodlog1 Ip person/

selection= cp b;

**Run**;

**Question3 Selection the best model via Stepwise**

title"Model selection via stepwise";

**proc** **reg** data=projectnew;

model water=day temp temp2 prodlog prodlog1 person/

selection=stepwise;

**run**;\*Check assumptions: residuals output";

**proc** **reg** data=projectnew;

model water=day temp prodlog;

output out=acheck r=resid;

**run**;

symbol1 v=circle i=sm80;

**Question4 Diagnostics: Assumption Check**

\*Check assumptions: Linearity Check";

axis1 label=(angle=**90**'water');

axis2 label=('day');

axis3 label=('temp');

axis4 label=('prodlog');

title"Linearity Check: Water vs day";

**proc** **sort** data=acheck; by day;

**proc** **gplot** data=acheck;

plot water\*day/vaxis=axis1 haxis=axis2;

**run**;

title"Linearity Check: Water vs temp";

**proc** **sort** data=acheck; by temp;

**proc** **gplot** data=acheck;

plot water\*temp/vaxis=axis1 haxis=axis3;

**run**;

title"Linearity Check: Water vs prodlog";

**proc** **sort** data=acheck; by prodlog;

**proc** **gplot** data=acheck;

plot water\*prodlog/vaxis=axis1 haxis=axis4;

**run**;

\*Check assumptions: residuals vs variables";

title"residuals vs X";

**proc** **gplot** data=acheck;

plot resid\*day/vref=**0**;

plot resid\*temp/vref=**0**;

plot resid\*prodlog/vref=**0**;

**Run**;

\*Check assumptions: normality test";

title"Normality Check";

**proc** **univariate** data=acheck plot normal;

var resid;

histogram resid/normal kernel(L=**2**);

**Run**;

**proc** **univariate** data=acheck plot normal;

var resid;

qqplot resid/normal(L=**1** mu=est sigma=est);

**run**;

\*Check assumptions: Independence";

title"Independence Check";

**data** scheck;

set acheck;

seq=\_n\_;

**run**;

symbol v=circle i=sm80;

**proc** **gplot** data=scheck;

plot resid\*seq;

**Run**;

**Question5 Diagnostic: Influence data points**

title"Check types of residuals";

**proc** **reg** data=acheck;

modelwater=day temp prodlog/

r influence;

**run**;

\*Get the graphs of diagnostics;

**ods graphics on;**  
**proc reg** data=acheck  
 plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);  
 model water=day temp prodlog;  
**run;**  
**ods graphics off;**

\*check multicollinearity;

\*VIF;

title"Measure of multicollinearity";

**proc** **reg** data=acheck;

model water=day temp prodlog/tol vif;

**Run**;

\*Partial Residual Plot;

title'Partial Residual plot';

**proc** **reg** data=acheck;

model water = day temp prodlog/r partial;

plot r.\*(day temp prodlog);

**run**;

**Question6 Summary & Prediction**

title"90% CI for the mean of response";

**proc** **reg** data=acheck;

model water=day temp prodlog/clm alpha=**0.1**;

**run**;

title"90% CI for regression coefficients";

**proc** **reg** data=acheck;

model water=day temp prodlog/clb alpha=**0.1**;

**run**;

title"90% PI of individual observations";

**proc** **reg** data=acheck;

model water=day temp prodlog/cli alpha=**0.1**;

**Run**;

**Support I:**

**\*Piecewise Prodlog**

data water;

input Ip tem prod day ppl W prodlog;

cards;

0 58.8 7107 21 129 3067 0.97180

0 65.2 6373 22 141 2828 1.04761

0 70.9 6796 22 153 2891 1.00463

0 77.4 9208 20 166 2994 0.71636

1 79.3 14792 25 193 3082 1.52940

1 81.0 14564 23 189 3898 1.48641

1 71.9 11964 20 175 3502 1.33549

1 63.9 13526 23 186 3060 1.26328

1 54.5 12656 20 190 3211 1.16702

1 39.5 14119 20 187 3286 1.39677

1 44.5 16691 22 195 3542 1.82986

0 43.6 14571 19 206 3125 0.38776

1 56.0 13619 22 198 3022 1.28543

1 64.7 14575 22 192 2922 1.48852

1 73.0 14556 21 191 3950 1.48486

1 78.9 18573 21 200 4488 2.05874

1 79.4 15618 22 200 3295 1.67128

;

\*the prodlog is the transformed variable with the prodlog axis changed according to Ip;

data sample;

set water;

if prodlog le 1

then p2=0;

if prodlog gt 1

then p2=prodlog-1;

run;

proc reg data=sample;

model W=prodlog p2;

output out=sample2 p=What;

run;

symbol1 v=circle i=none c=black;

symbol2 v=none i=join c=black;

proc sort data=sample2;by prodlog;

proc gplot data=sample2;

plot(W What)\*prodlog/overlay;

Run;

**Support II:**

**\*Remedial Measures: Weighted Regression**

\*For temperature;

**data** checkt;

set acheck;

absr=abs(resid);

sqrr=resid\*resid;

**run**;

title"transresidual vs temp";

**proc** **gplot** data=checkt;

plot(resid absr sqrr)\*temp;

**run**;

**proc** **reg** data=checkt;

model absr=temp;

output out=findtemp p=shat;

**run**;

**data** findtemp;

set findtemp;

wt=**1**/(shat\*shat);

**run**;

**proc** **reg** data=findtemp;

model water=temp/clb p;

weight wt;

output out=weighted p=predict;

**run**;

\*For prodlog';

**data** checkprod;

set acheck;

absr=abs(resid);

sqrr=resid\*resid;

**run**;

title"transresidual vs prodlog";

**proc** **gplot** data=checkprod;

plot(resid absr sqrr)\*prodlog;

**run**;

**proc** **reg** data=checkprod;

model absr=prodlog;

output out=findprod p=shat2;

**run**;

**data** findprod;

set findprod;

wt=**1**/(shat2\*shat2);

**run**;

**proc** **reg** data=findprod;

model water=prodlog/clb p;

weight wt;

output out=weighted2 p=predict2;

**run**;