Predictive Modelling with Linear Regression

Full name: **Kunal Sevak** Student number: **8817782** Course number: **1498**

Assignment

The goal of this project is for you to perform data analysis, predictive modeling, and diagnostics on a data set that represents Total Rice Production in the Country of Indonesia.

The statistical method used is linear regression and multiple linear regression.

Key Elements of the Project:

1. Data Set Analysis:

The Data Set that we have selected is Total Rice Production in Indonesia.

The below details are of our total Numeric Variables or our predictors.

Description of Rice Production Data Set			
Sr No	**Variables**	**Description**	**Remarks**
1	Size	Size or Area of the entire field	Predictors
2	Seed	Total Seeds procured for Rice Production	Predictors
3	Urea	Total Urea Organic Fertilizer procured	Predictors
4	Phosphate	Total Phosphate procured as Fertilizer	Predictors
5	Pesticide	Total Pesticide Procured	Predictors
6	Pseed	Total Seeds used for Cultivation	Predictors
7	Purea	Total Urea Used for Cultivation	Predictors
8	Pphosphate	Total Phosphate Used for Cultivation	Predictors
9	Hired Labor	Total Cost of Hired Labor for the Production	Predictors
10	Family Labor	Total Cost of Family Labour	Predictors
11	Total Labor	Total Cost of the Labor	Predictors
12	Wage	Total Cost of Wages	Predictors
13	Gross Output	Gross Output of Rice Production	Predictors
14	Net Output	Net Out Put of Rice Production	Response Variable
15	Price	Total Price	Response Variable

- We have selected Net Output and Price as our Response Variables as described in the above table because it is a performance parameter for total Rice production in the country.
- Prediction of these response variables is essential as it will give a clear indication for forecasting Price and Net Output based on all the predictors.
- There are more than 1000 Rows of entries for the complete Data Set, which describes key variables for predicting our response variables.
- Our aim is to find the best prediction for the Price and Net Output of Rice Production based on our predictors mentioned above by analyzing the relationship of the Response variable with each predictor.

Data Set and Libraries Loading in R Studio:

```
library(ggplot2)
library(tidyverse)
library(ISLR2)
library(caret)
library(stargazer)
library(lmtest)
library(leaps)
library(readr)

RiceData <- read_csv("C:/Users/This Pc/Desktop/Conestoga Sem 1/Multivariate Statistics STAT8030/Project 1 - Linear Regression/Data Set/Final Data Set/Production of Rice in Indonesia.csv")
```

Summary of Entire Data Set:

```
> summary(RiceData)
      ...1             id              size           status
 Min.   :   1.0   Min.   :101001   Min.   :0.0100   Length:1026
 1st Qu.: 257.2   1st Qu.:209250   1st Qu.:0.1430   Class :character
 Median : 513.5   Median :401037   Median :0.2860   Mode  :character
 Mean   : 513.5   Mean   :374954   Mean   :0.4316
 3rd Qu.: 769.8   3rd Qu.:504162   3rd Qu.:0.5000
 Max.   :1026.0   Max.   :609245   Max.   :5.3220
  varieties            bimas                seed              urea
 Length:1026        Length:1026        Min.   :   1.00   Min.   :   1.00
 Class :character   Class :character   1st Qu.:   5.00   1st Qu.:  25.00
 Mode  :character   Mode  :character   Median :  10.00   Median :  60.00
                                       Mean   :  18.21   Mean   :  95.44
                                       3rd Qu.:  20.00   3rd Qu.: 100.00
                                       Max.   :1250.00   Max.   :1250.00
   phosphate        pesticide         pseed            purea
 Min.   :  0.00   Min.   :    0   Min.   : 40.0   Min.   : 50.00
 1st Qu.:  8.00   1st Qu.:    0   1st Qu.: 70.0   1st Qu.: 70.00
 Median : 20.00   Median :    0   Median : 81.0   Median : 80.00
 Mean   : 33.73   Mean   :  595   Mean   :112.1   Mean   : 78.98
 3rd Qu.: 50.00   3rd Qu.:  265   3rd Qu.:150.0   3rd Qu.: 85.00
 Max.   :700.00   Max.   :62600   Max.   :375.0   Max.   :100.00
    pphosph         hiredlabor      famlabor         totlabor
 Min.   : 60.00   Min.   :   1   Min.   :   1.0   Min.   :  17.0
 1st Qu.: 70.00   1st Qu.:  36   1st Qu.:  69.0   1st Qu.: 144.0
 Median : 80.00   Median : 112   Median : 111.0   Median : 252.0
 Mean   : 79.57   Mean   : 237   Mean   : 151.5   Mean   : 388.4
 3rd Qu.: 85.00   3rd Qu.: 260   3rd Qu.: 185.0   3rd Qu.: 435.0
 Max.   :120.00   Max.   :4536   Max.   :1526.0   Max.   :4774.0
      wage            goutput           noutput          price
 Min.   : 30.00   Min.   :   42.0   Min.   :   42   Min.   : 50.00
 1st Qu.: 49.38   1st Qu.:  420.0   1st Qu.:  380   1st Qu.: 60.50
 Median : 57.14   Median :  886.5   Median :  800   Median : 75.00
 Mean   : 80.42   Mean   : 1405.2   Mean   : 1241   Mean   : 90.96
 3rd Qu.:128.75   3rd Qu.: 1606.0   3rd Qu.: 1444   3rd Qu.:120.00
 Max.   :175.35   Max.   :20960.0   Max.   :17610   Max.   :190.00
    region
 Length:1026
 Class :character
 Mode  :character
```

From the above summary, we get the idea regarding every variable present in the data set, and also it gives an accurate 5-point summary of each variable which can be used by us for future references while making our predictions.

2. Initial Modelling:

Initial modeling is based on a linear relationship between predictors and response variables.

For checking the linear relationship, we take the correlation value between each predictor and our response variable.

The correlation value is always between -1 and 1, -1 indicating a negative linear relationship, 0 indicating no linear relationship, and 1 indicating a positive linear relationship.

- **Linear Relationship between Price and all the Predictors:**

```
> cor(RiceData$price,RiceData$noutput)
[1] 0.09852095
> # Positive Linear Relation between Price of Rice and Net Output of Rice = 0.09852095
> cor(RiceData$price,RiceData$size)
[1] -0.01223299
> # Negative Linear Relationship between price of Rice and Size of field = -0.01223299
> cor(RiceData$price,RiceData$seed)
[1] -0.02642857
> # -0.02642857
> cor(RiceData$price,RiceData$urea)
[1] 0.07534964
> # 0.07534964
> cor(RiceData$price,RiceData$phosphate)
[1] 0.1899533
> # 0.1899533
> cor(RiceData$price,RiceData$pesticide)
[1] 0.1061548
> #0.1061548
> cor(RiceData$price,RiceData$pseed)
[1] 0.6689168
> #0.6689168
> cor(RiceData$price,RiceData$pphosph)
[1] 0.6878633
> #0.687863
> cor(RiceData$price,RiceData$purea)
[1] 0.6849733
> #0.6849733
> cor(RiceData$price,RiceData$hiredlabor)
[1] -0.003335649
> #-0.003335649
> cor(RiceData$price,RiceData$famlabor)
[1] 0.109077
> # 0.109077
> cor(RiceData$price,RiceData$totlabor)
[1] 0.03012847
> #0.03012847
> cor(RiceData$price,RiceData$wage)
[1] 0.8593039
> #0.8593039
> cor(RiceData$price,RiceData$goutput)
[1] 0.09119443
> #0.09119443
> cor(RiceData$price,RiceData$noutput)
[1] 0.09852095
>
```

Here for the above table, we are highlighting 2 predictors that have the best linear relationship with our response variable Price which we will use to do further linear modeling for predictions.

- **Linear Relationship between Our Response Variable Price with Predictors:**

Linear Relationship Between Our Response Variable Price with Predictors			
Sr No	**Predictors**	**Correlation Value with Response Variable Price**	**Remarks**
1	Size	-0.01223299	Negative relationship but closer to 0
2	Seed	-0.02642857	Negative relationship but closer to 0
3	Urea	0.07534964	Positive Relationship but closer to 0
4	Phosphate	0.1899533	Positive Linear Relationship
5	Pesticide	0.1061548	Positive Linear Relationship
6	Pseed	0.6689168	Good Positive Linear Relationship
7	Pphosphate	0.687863	Good Positive Linear Relationship
8	Purea	0.6849733	Good Positive Linear Relationship
9	Hired Labor	-0.003335649	Negative relationship but closer to 0
10	Family Labor	0.109077	Positive Relationship but closer to 0
11	Total Labor	0.03012847	Positive Relationship but closer to 0
12	Wage	0.8593039	Good Positive Linear Relationship
13	Gross Output	0.09119443	Positive Relationship but closer to 0
14	Net Output	0.09852095	Positive Relationship but closer to 0

- **Linear Relationship between Net Output and all the Predictors:**

```
> # Correlations with Net Out Put and Other Variables
>
> cor(RiceData$noutput,RiceData$size)
[1] 0.8915277
> #0.8915277
> cor(RiceData$noutput,RiceData$seed)
[1] 0.5475009
> #0.5475009
> cor(RiceData$noutput,RiceData$urea)
[1] 0.8134663
> #0.8134663
> cor(RiceData$noutput,RiceData$phosphate)
[1] 0.7370739
> #0.7370739
> cor(RiceData$noutput,RiceData$pesticide)
[1] 0.3951422
> # 0.3951422
> cor(RiceData$noutput,RiceData$pseed)
[1] 0.1532252
> # 0.1532252
> cor(RiceData$noutput,RiceData$pphosph)
[1] 0.0217065
> # 0.0217065
> cor(RiceData$noutput,RiceData$purea)
[1] 0.02984448
> # 0.02984448
> cor(RiceData$noutput,RiceData$hiredlabor)
[1] 0.8511969
> # 0.8511969
> cor(RiceData$noutput,RiceData$famlabor)
[1] 0.4115718
> # 0.4115718
> cor(RiceData$noutput,RiceData$totlabor)
[1] 0.8681
> # 0.8681
> cor(RiceData$noutput,RiceData$wage)
[1] 0.1714056
> # 0.17174056
> cor(RiceData$noutput,RiceData$goutput)
[1] 0.9988217
> # 0.9988217
> cor(RiceData$noutput,RiceData$price)
[1] 0.09852095
> # 0.09852095
```

Here for the above table, we are highlighting 3 predictors that have the best linear relationship with our response variable Net Output which we will use to do further linear modeling for predictions.

- **Linear Relationship between Our Response Variable Net Output with Predictors:**

Linear Relationship Between Our Response Variable Net Output with Predictors			
Sr No	**Predictors**	**Correlation Value with Response Variable Price**	**Remarks**
1	Size	0.8915277	Good Positive Linear Relationship
2	Seed	0.5475009	Positive Linear Relationship
3	Urea	0.8134663	Good Positive Linear Relationship
4	Phosphate	0.7370739	Positive Linear Relationship
5	Pesticide	0.3951422	Positive Linear Relationship
6	Pseed	0.1532252	Positive Linear Relationship
7	Pphosphate	0.0217065	Positive Relationship but closer to 0
8	Purea	0.02984448	Positive Relationship but closer to 0
9	Hired Labor	0.8511969	Positive Relationship but closer to 0
10	Family Labor	0.4115718	Positive Linear Relationship
11	Total Labor	0.8681	Good Positive Linear Relationship
12	Wage	0.17174056	Positive Linear Relationship
13	Gross Output	0.9988217	Good Positive Linear Relationship
14	Price	0.09852095	Positive Relationship but closer to 0

3. Diagnostics on Linear Models Prepared based on the above details:

- We will highlight all the Models for our Response Variable Price and Linear Models their Coefficient value, Graphs, Residual and Fitted Values, Stargazer Summary and RMSE, R Squared Values, Adjusted R Square values, etc.
- In addition, we will plot necessary graphs to find a linear relationship between the response variable and the predictor/s.
- Also, we will be using the library stargazer and its function stargazer to find various statistical relationships and summary between variables.
- We will plot residual and fitted values to analyze the patterns.

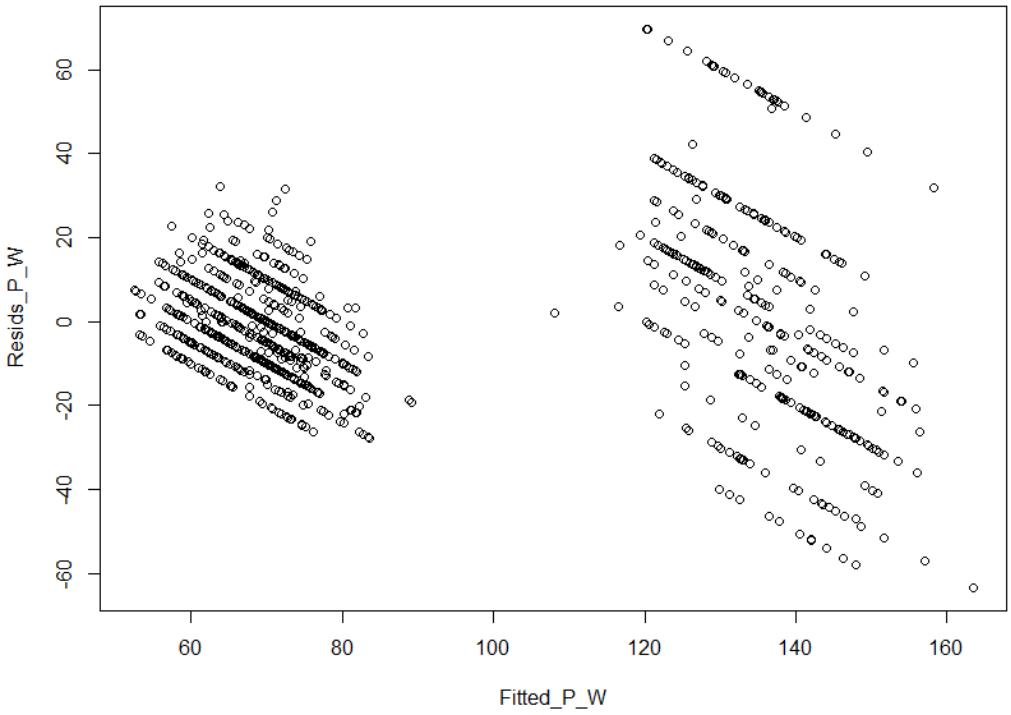
➢ Models and Calculations for Response Variable Price and Net Output.

➢ Price with all the predictors

1. Price and Wage Relationship:

```
> price_regression_wage <- lm(price ~ wage, data=RiceData)
> coef(price_regression_wage)
(Intercept)        wage
 29.5421133   0.7637004
> RiceData %>% ggplot(aes(x = price, y = wage)) + geom_point() + geom_smooth(method="lm", se = FALSE)
`geom_smooth()` using formula 'y ~ x'
> Resids_P_W <- price_regression_wage$residuals
> Fitted_P_W <- price_regression_wage$fitted.values
> RiceData %>% ggplot(aes(x = wage, y = price)) + geom_point() + geom_smooth(method="lm", se = FALSE)
`geom_smooth()` using formula 'y ~ x'
>
```


price
160
120
80
50
100
150
wage

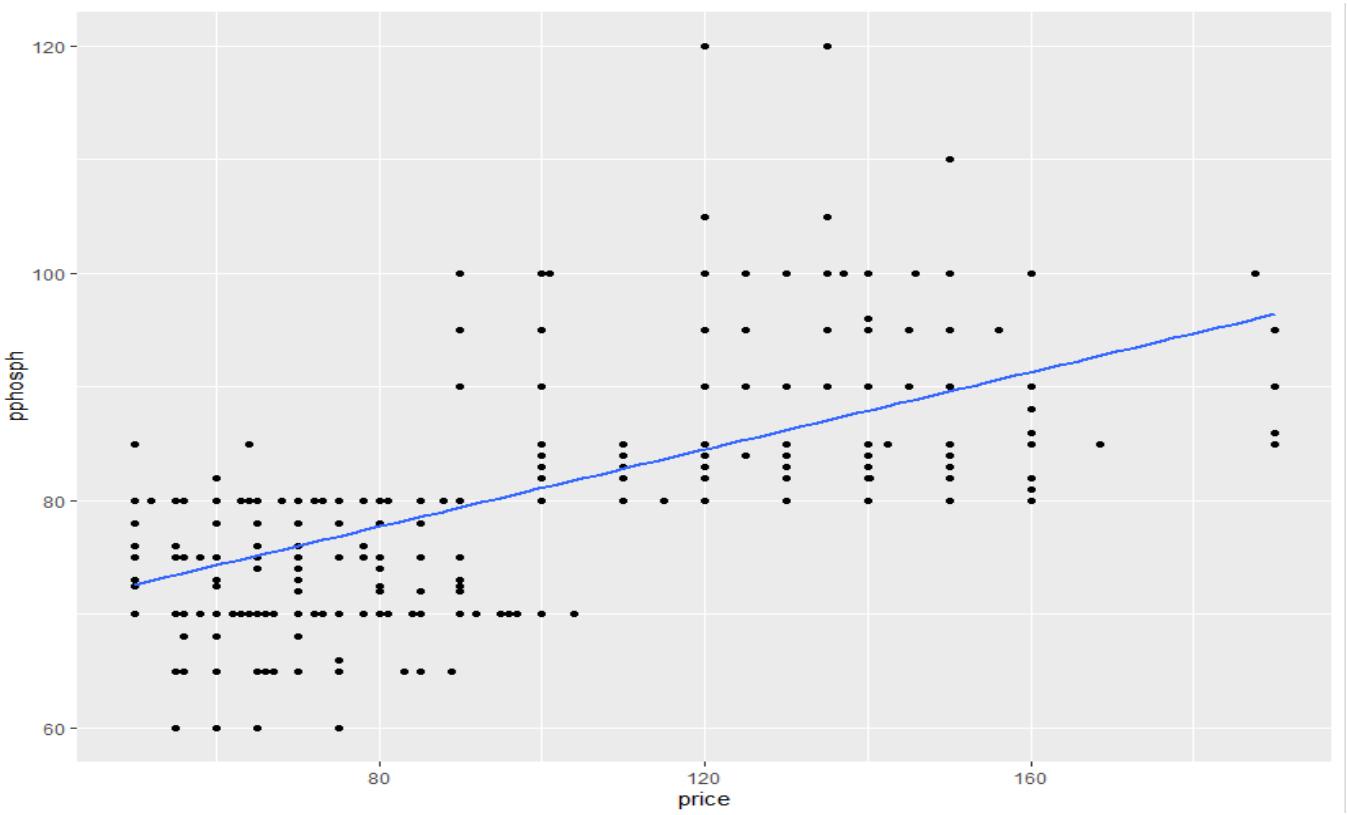

Resids_P_W
60
40
20
0
-20
-40
-60
60
80
100
120
140
160
Fitted_P_W

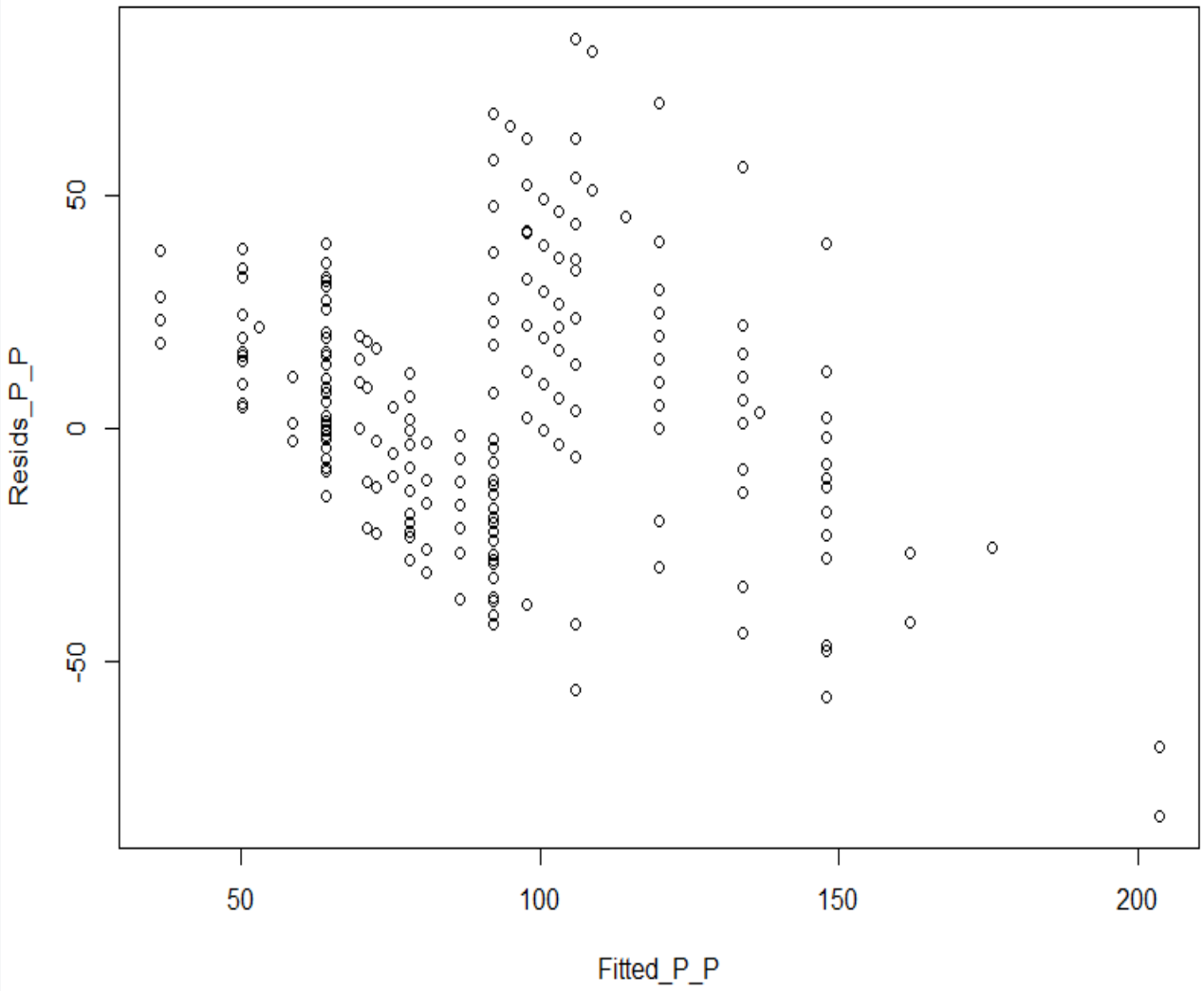
```
================================================
                        Dependent variable:
                    ---------------------------
                                price
------------------------------------------------
wage                          0.764***
                               (0.014)

Constant                      29.542***
                               (1.290)

------------------------------------------------
Observations                    1,026
R2                              0.738
Adjusted R2                     0.738
Residual Std. Error      19.187 (df = 1024)
F Statistic          2,890.421*** (df = 1; 1024)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> |
```

```
> P_W_CVModel <- train(
+   form = price ~ wage,
+   data=RiceData,
+   method = "lm",
+   trControl = trainControl(method = "cv", number = 10)
+ )
> P_W_CVModel
Linear Regression

1026 samples
   1 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 923, 923, 925, 922, 924, 924, ...
Resampling results:

  RMSE      Rsquared   MAE
  19.04809  0.7438081  14.18327

Tuning parameter 'intercept' was held constant at a value of TRUE
> |
```

2. Price and Pphosphate:

```
> price_regression_pphos <- lm(price ~ pphosph, data = RiceData)
> coef(price_regression_pphos)
(Intercept)     pphosph
-130.370791    2.781683
> RiceData %>% ggplot(aes(x = price, y = pphosph)) + geom_point() + geom_smooth(method="lm", se = FALSE)
`geom_smooth()` using formula 'y ~ x'
>
```



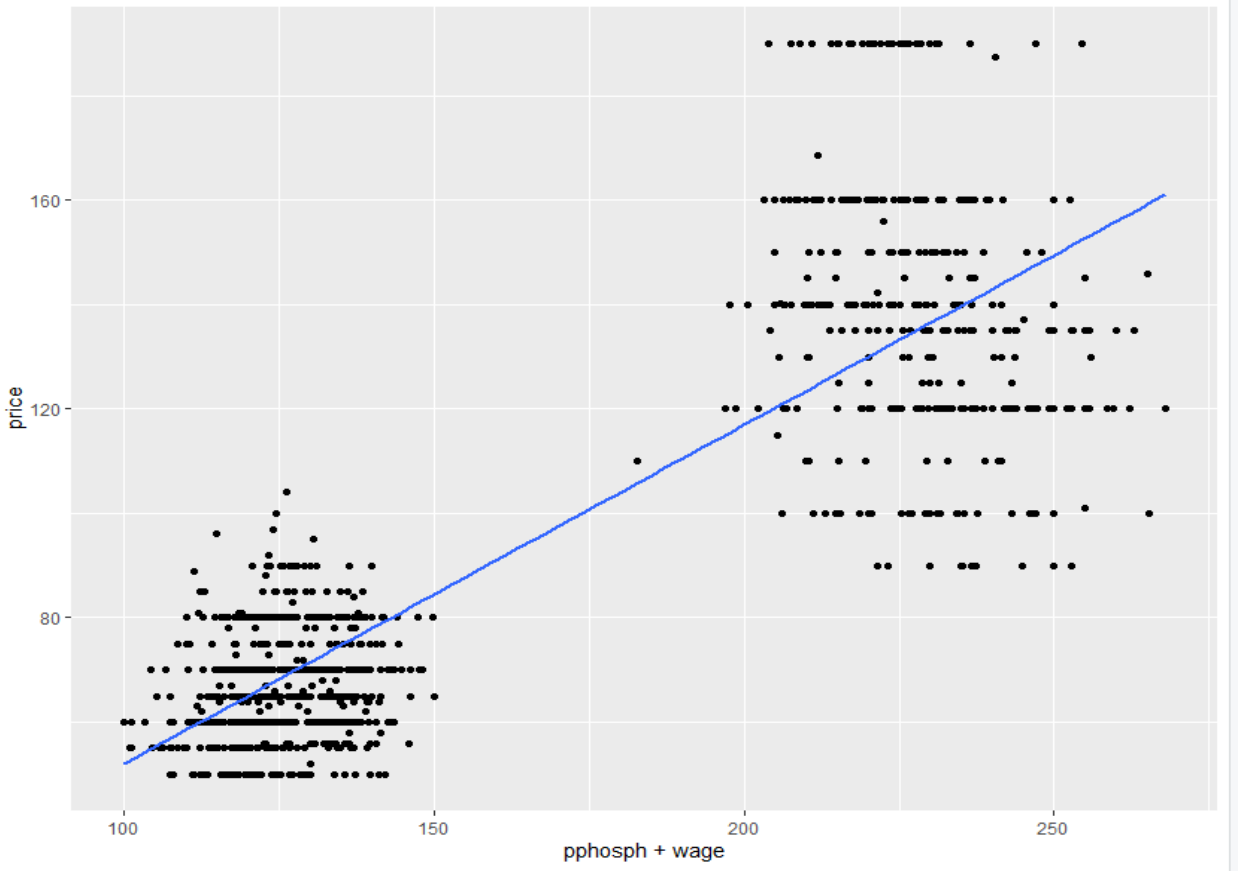


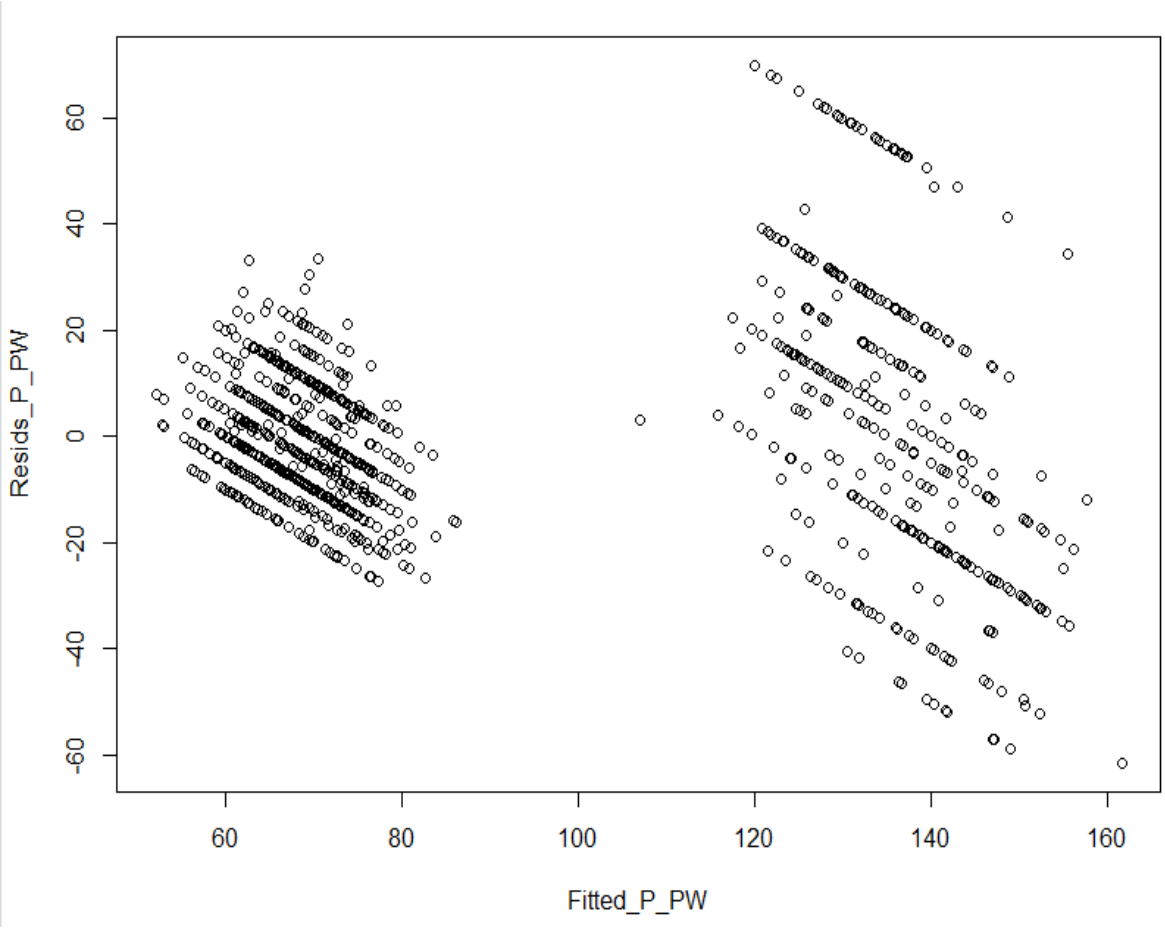

```
================================================
                         Dependent variable:
                     ---------------------------
                                price
------------------------------------------------
pphosph                        2.782***
                               (0.092)

Constant                      -130.371***
                               (7.348)

------------------------------------------------
Observations                     1,026
R2                               0.473
Adjusted R2                      0.473
Residual Std. Error        27.229 (df = 1024)
F Statistic             919.649*** (df = 1; 1024)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> |
```

```
> P_P_CVModel <- train(
+   form = price ~ pphosph,
+   data=RiceData,
+   method = "lm",
+   trControl = trainControl(method = "cv", number = 10)
+ )
> P_P_CVModel
Linear Regression

1026 samples
   1 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 923, 922, 923, 924, 922, 924, ...
Resampling results:

  RMSE      Rsquared   MAE
  27.20919  0.4804039  20.66271

Tuning parameter 'intercept' was held constant at a value of TRUE
>
```

3. Price with Wage and Pphosphate:

```
> price_regression_pphos_Wage <- lm(price ~ wage + pphosph, data = RiceData)
> coef(price_regression_pphos_Wage)
(Intercept)        wage     pphosph
  7.4921236   0.7080163   0.3334051
> RiceData %>% ggplot(aes(x = price, y = pphosph + wage)) + geom_point() + geom_smooth(method="lm", se = FALSE)
`geom_smooth()` using formula 'y ~ x'
> RiceData %>% ggplot(aes(x = pphosph + wage, y = price)) + geom_point() + geom_smooth(method="lm", se = FALSE)
`geom_smooth()` using formula 'y ~ x'
>
```


price
160
120
80
100
150
200
250
pphosph + wage

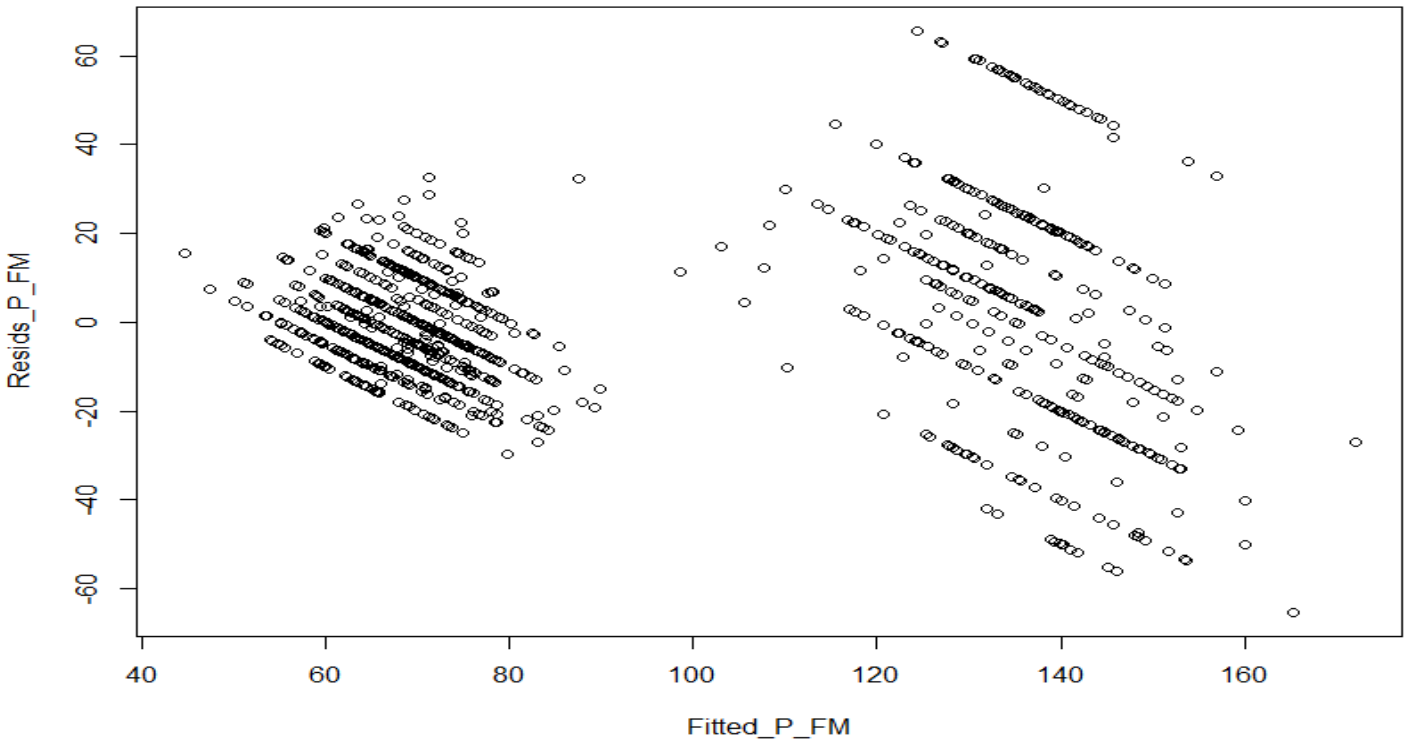

Resids_P_PW
60
40
20
0
-20
-40
-60
60
80
100
120
140
160
Fitted_P_PW

```
> stargazer(price_regression_pphos_Wage, type ="text")

================================================
                        Dependent variable:
                    ----------------------------
                               price
------------------------------------------------
wage                          0.708***
                               (0.022)

pphosph                       0.333***
                               (0.099)

Constant                       7.492
                              (6.668)

------------------------------------------------
Observations                   1,026
R2                             0.741
Adjusted R2                    0.741
Residual Std. Error      19.090 (df = 1023)
F Statistic          1,465.501*** (df = 2; 1023)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
> P_WP_CVModel <- train(
+   form = price ~ wage + pphosph,
+   data=RiceData,
+   method = "lm",
+   trControl = trainControl(method = "cv", number = 10)
+ )
> P_WP_CVModel
Linear Regression

1026 samples
   2 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 923, 924, 924, 922, 924, 924, ...
Resampling results:

  RMSE      Rsquared   MAE
  19.05313  0.7447399  14.05486

Tuning parameter 'intercept' was held constant at a value of TRUE
>
```

4. Price with Full Model:

```
> price_regression_fullModel <- lm(price ~ size + seed + urea + phosphate + pesticide + pseed + purea + pphosph + hiredlabor +
mlabor + totlabor + wage + goutput + noutput, data = RiceData)
> coef(price_regression_fullModel)
 (Intercept)         size         seed         urea    phosphate    pesticide        pseed        purea      pphosph
-2.385464449  2.363876111 -0.004179025  0.005798494 -0.044094972 -0.000789325  0.074414309  0.726631876 -0.313359427
  hiredlabor     famlabor     totlabor         wage      goutput      noutput
 0.131104417  0.129589448 -0.125677397  0.645420700 -0.034305256  0.038216661
>
```



	Dependent variable:
	price
size	2.364
	(3.540)
seed	-0.004
	(0.017)
urea	0.006
	(0.010)
phosphate	-0.044**
	(0.020)
pesticide	-0.001***
	(0.0002)
pseed	0.074***
	(0.014)
purea	0.727***
	(0.258)
pphosph	-0.313
	(0.249)
hiredlabor	0.131
	(0.125)
famlabor	0.130
	(0.126)
totlabor	-0.126
	(0.125)
wage	0.645***
	(0.029)
goutput	-0.034***
	(0.007)
noutput	0.038***
	(0.008)
Constant	-2.385
	(7.223)
Observations	1,026
R2	0.761
Adjusted R2	0.758
Residual Std. Error	18.438 (df = 1011)
F Statistic	230.563*** (df = 14; 1011)

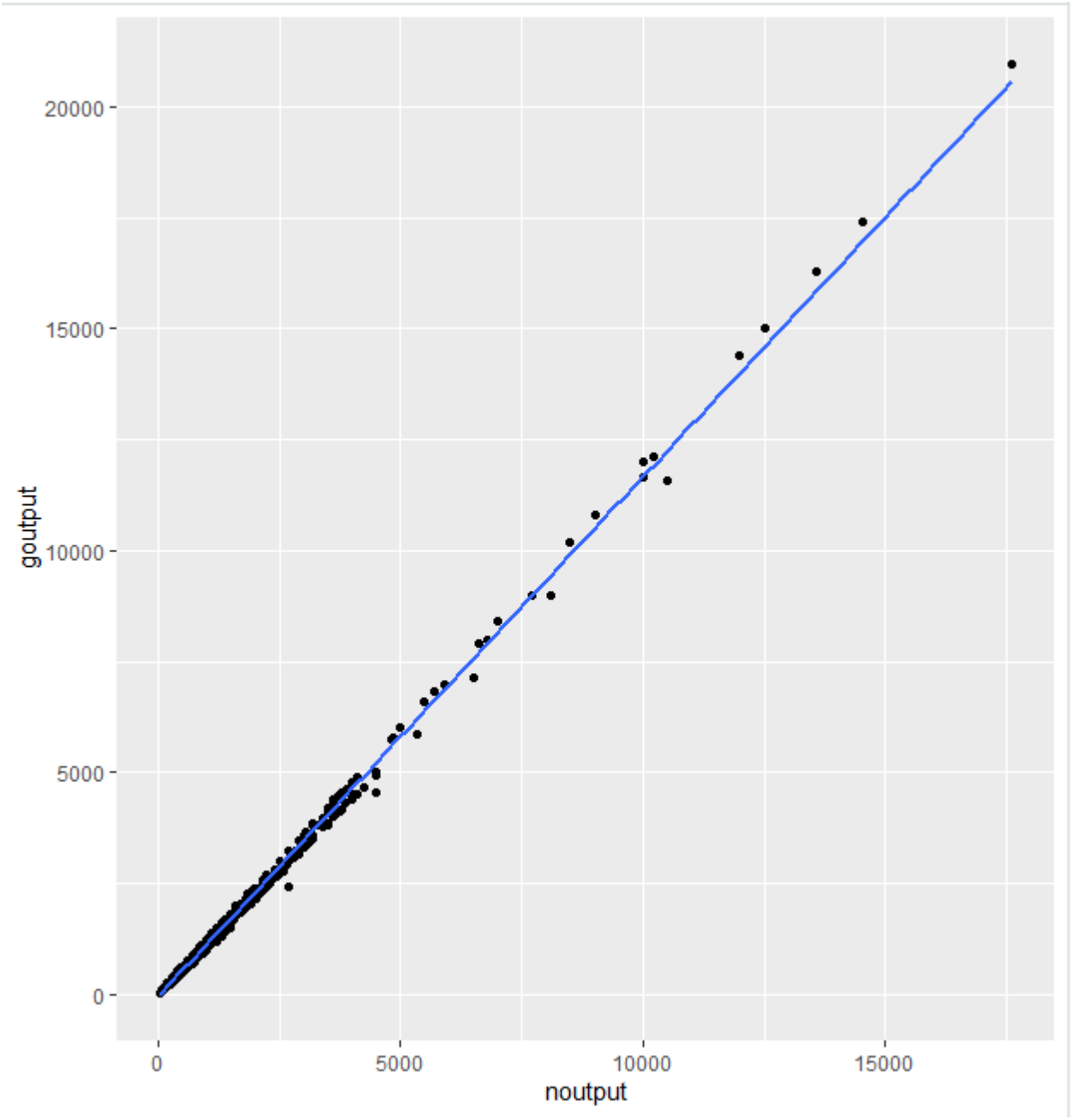
```
> P_FM_CVModel <- train(
+   form = price ~ .,
+   data=RiceData,
+   method = "lm",
+   trControl = trainControl(method = "cv", number = 10)
+ )
> P_FM_CVModel
Linear Regression

1026 samples
  20 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 923, 922, 925, 924, 922, 923, ...
Resampling results:

  RMSE      Rsquared   MAE
  16.75573  0.8024548  12.74964

Tuning parameter 'intercept' was held constant at a value of TRUE
```

From the above models we can consider Price can be best predicted when considering the full model, but we will perform some more technical parameters before we select the best model.

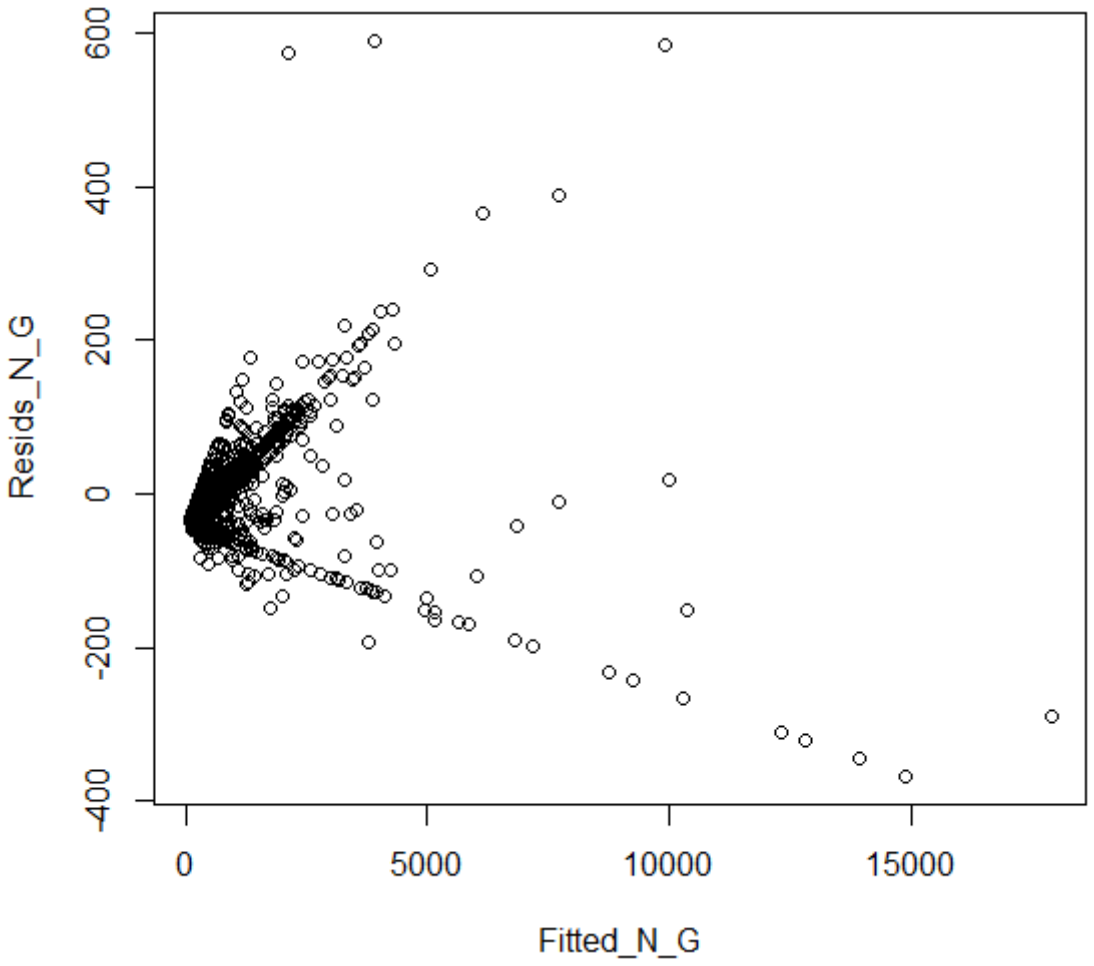
➢ Models and Calculations for Response Variable Net Output:

1. Net Output and Gross Output:

```
> Netoutput_regression_Grossoutput <- lm(noutput ~ goutput , data= RiceData)
> coef(Netoutput_regression_Grossoutput)
(Intercept)     goutput
 43.9262983   0.8518518
```



The above graph depicts that the linear relationship between Net Output and Gross Output is the best one.

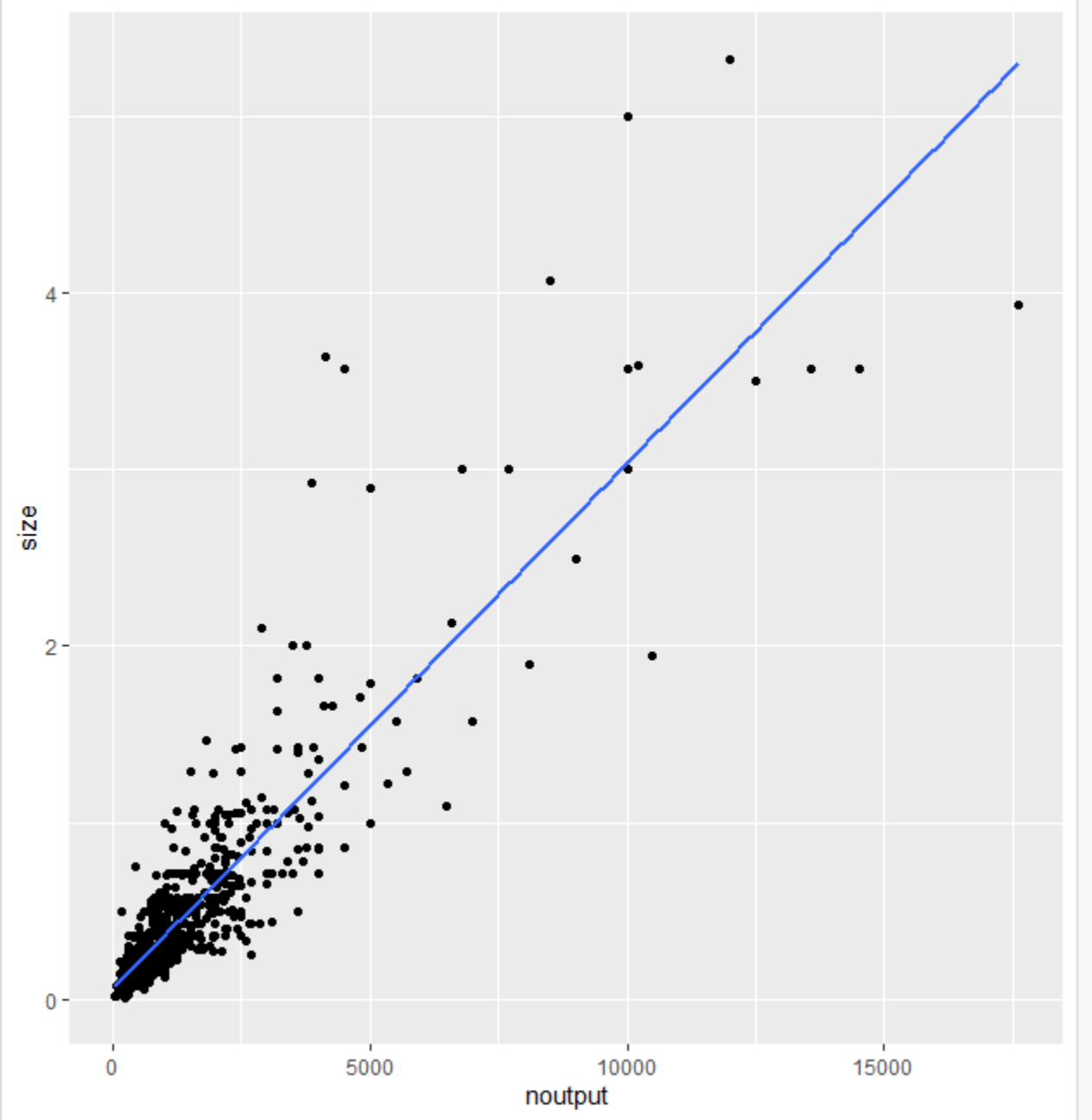


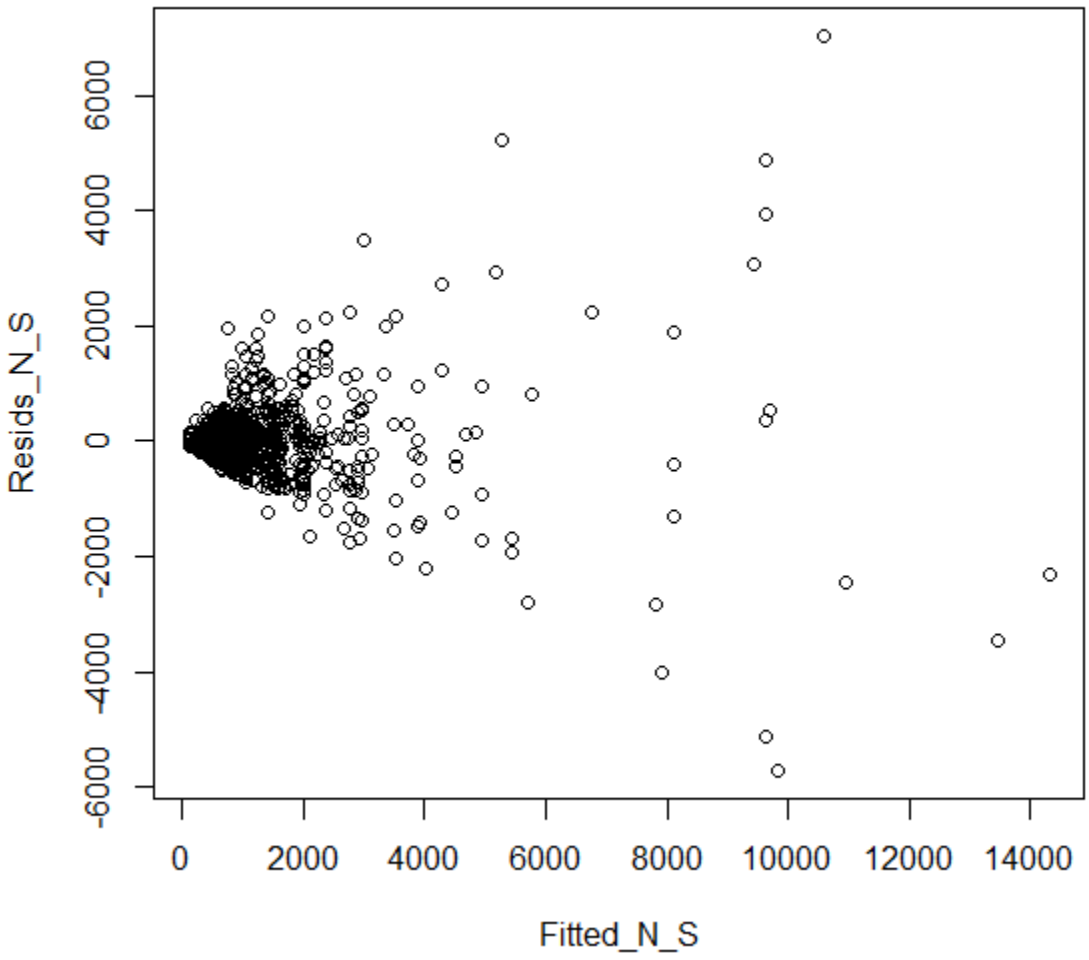

```
> stargazer(Netoutput_regression_Grossoutput, type ="text

================================================
                         Dependent variable:
                     ---------------------------
                                noutput
------------------------------------------------
goutput                        0.852***
                                (0.001)

Constant                       43.926***
                                (3.078)

------------------------------------------------
Observations                      1,026
R2                                0.998
Adjusted R2                       0.998
Residual Std. Error         79.580 (df = 1024)
F Statistic           433,747.600*** (df = 1; 1024)
================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
> |
```

```
Linear Regression

1026 samples
   1 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 923, 924, 924, 922, 924, 923, ...
Resampling results:

  RMSE      Rsquared   MAE
  79.39765  0.9974233  50.91422

Tuning parameter 'intercept' was held constant at a value of TRUE
> |
```

2. Net Output and Size:

```
> Netoutput_regression_size <- lm(noutput ~ size, data=RiceData)
> coef(Netoutput_regression_size)
(Intercept)        size
   87.53125  2672.37165
> |
```


size
noutput
0
2
4
0
5000
10000
15000

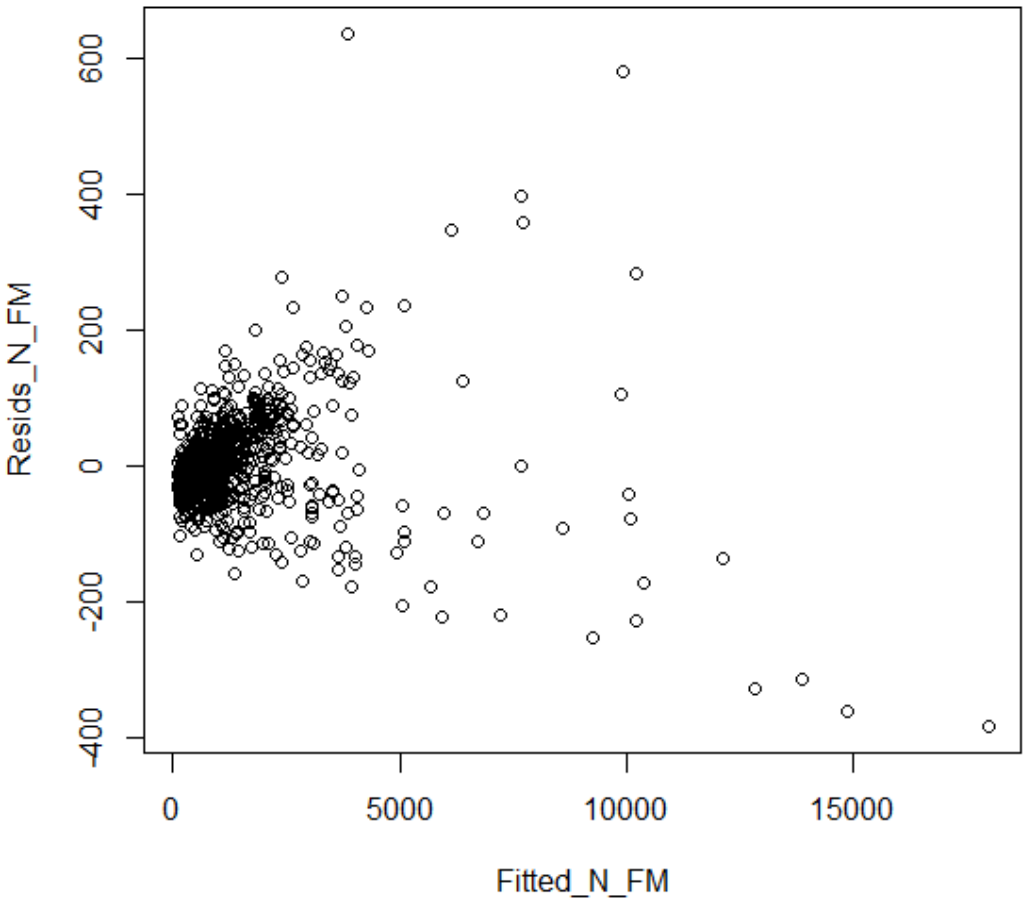



```
================================================
                        Dependent variable:
                    ---------------------------
                               noutput
------------------------------------------------
size                         2,672.372***
                               (42.430)

Constant                       87.531***
                               (29.548)

------------------------------------------------
Observations                     1,026
R2                               0.795
Adjusted R2                      0.795
Residual Std. Error       742.766 (df = 1024)
F Statistic           3,966.780*** (df = 1; 1024)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> |
```

```
Linear Regression

1026 samples
   1 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 925, 923, 923, 924, 924, 923, ...
Resampling results:

  RMSE      Rsquared   MAE
  722.7337  0.7978275  384.4727

Tuning parameter 'intercept' was held constant at a value of TRUE
> |
```

3. Net Output with all the predictors:

```
> NetOutput_regression_fullModel <- lm(noutput ~ size + seed + urea + phosphate + pesticide + pseed + purea + pphosph + hiredlabor + famlabor + totlabor
 wage + goutput + price, data = RiceData)
> coef(NetOutput_regression_fullModel)
  (Intercept)          size          seed          urea     phosphate     pesticide         pseed         purea       pphosph    hiredlabor
-0.884488999 -63.749211039  -0.065612253   0.114621020   0.361586508   0.005216177  -0.440947854   0.314058211   0.109525188   2.210909561
     famlabor      totlabor          wage       goutput         price
 2.226267116  -2.161523982   0.026369319   0.843065190   0.569408713
> |
```



```
                          Dependent variable:
                      ---------------------------
                                 noutput
------------------------------------------------
size                          -63.749***
                               (13.518)

seed                           -0.066
                               (0.066)

urea                          0.115***
                               (0.038)

phosphate                     0.362***
                               (0.076)

pesticide                     0.005***
                               (0.001)

pseed                         -0.441***
                               (0.053)

purea                          0.314
                               (0.999)

pphosph                        0.110
                               (0.960)

hiredlabor                    2.211***
                               (0.479)

famlabor                      2.226***
                               (0.481)

totlabor                      -2.162***
                               (0.480)

wage                           0.026
                               (0.138)

goutput                       0.843***
                               (0.003)

price                         0.569***
                               (0.120)

Constant                       -0.884
                              (27.882)

------------------------------------------------
Observations                    1,026
R2                              0.998
Adjusted R2                     0.998
Residual Std. Error       71.170 (df = 1011)
F Statistic         38,756.550*** (df = 14; 1011)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> |
```

```
Linear Regression

1026 samples
  20 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 925, 924, 923, 923, 923, 922, ...
Resampling results:

  RMSE      Rsquared  MAE
  75.52169  0.997846  43.30572

Tuning parameter 'intercept' was held constant at a value of TRUE
```

According to the above observations with all the models, we can understand that our 2 models – One with Gross Output and the other with All the variables are very close to each other, but we will use further methods to select the best model.

4. Model Selection:

We have selected the Best Model based on various analyses such as getting P Values, RMSE, R Squared, MAE Values, Getting Residual values, and fitted values.

According to our analysis, we have the 2 Best Linear Models for our Response Variables Price and Net Output.

The R function regsubsets() [leaps package] can be used to identify different best models of different sizes by using Adjusted R Squared Values.

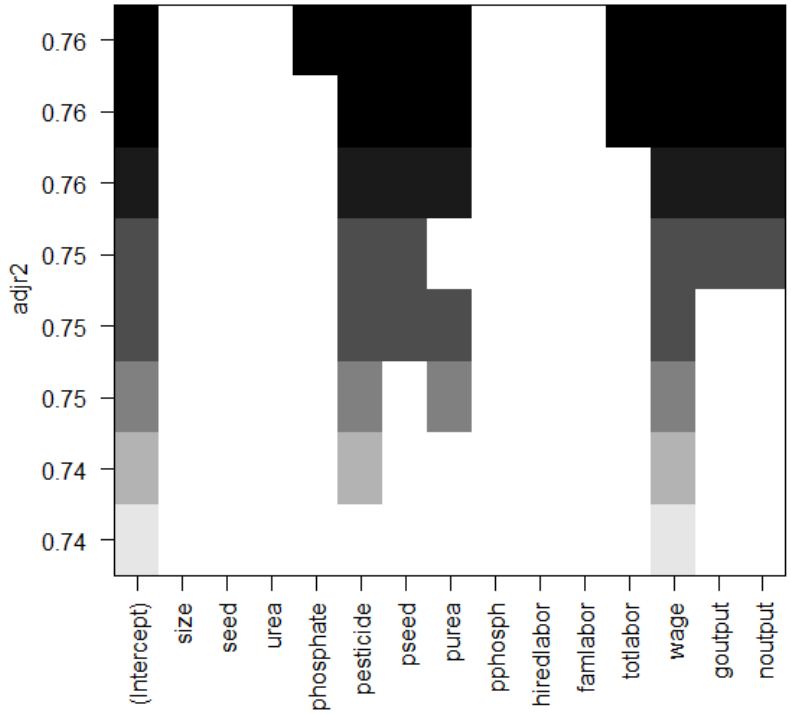
The adjusted R2 represents the proportion of variation, in the outcome, that is explained by the variation in the predictor's values. The higher the adjusted R2, the better the model.

As we can check from the results that there are lower the value of RMSE and MAE from the above models the below model is much better for making accurate predictions.

In addition, The Values of R Squared are also closer to 1, indicating that our results are good for making predictions.

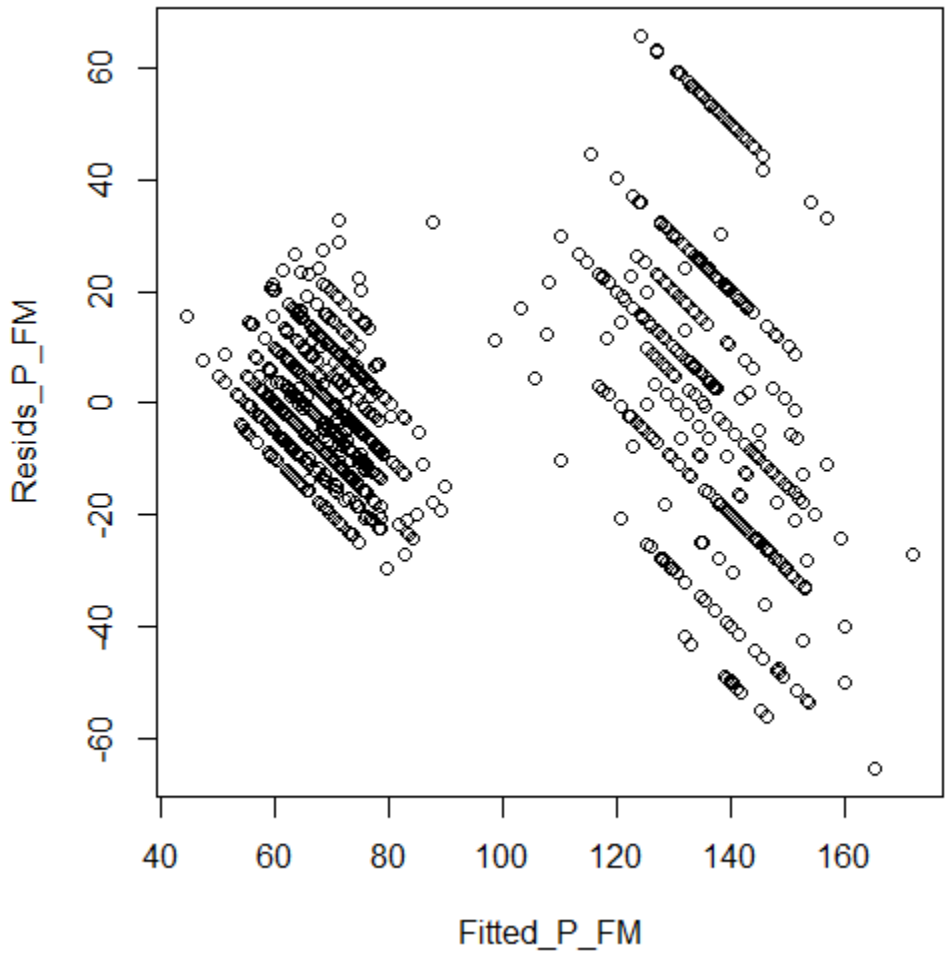
Output with Best Model using Reg Subsets and Adjusted R Squared Values for Response Variable Price and Net Output:

Price Best Model:



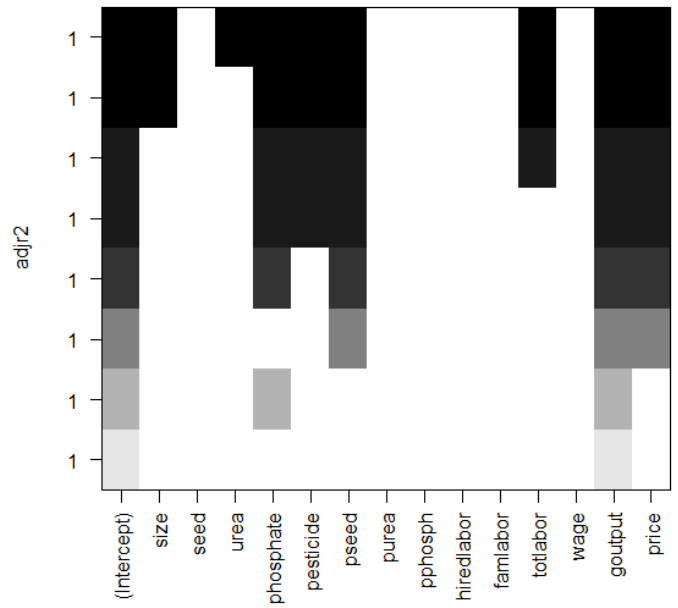

```
> price_regression_BM <- lm(price ~ urea + phosphate + pesticide + pseed + purea + totlabor + wage + goutput + noutput, data = RiceData)
> coef(price_regression_BM)
  (Intercept)          urea     phosphate     pesticide         pseed         purea      totlabor          wage       goutput       noutput
-3.0207740415  0.0082348584 -0.0449130824 -0.0008104783  0.0743915444  0.4267551349  0.0058950786  0.6359887566 -0.0343822507  0.0387017466
> |
```

```
================================================
                         Dependent variable:
                     ---------------------------
                                price
------------------------------------------------
urea                             0.008
                                (0.008)

phosphate                       -0.045**
                                 (0.019)

pesticide                       -0.001***
                                (0.0002)

pseed                           0.074***
                                 (0.014)

purea                           0.427***
                                 (0.103)

totlabor                         0.006**
                                 (0.003)

wage                            0.636***
                                 (0.028)

goutput                         -0.034***
                                 (0.007)

noutput                         0.039***
                                 (0.008)

Constant                         -3.021
                                 (7.104)

------------------------------------------------
Observations                      1,026
R2                                0.761
Adjusted R2                       0.759
Residual Std. Error      18.424 (df = 1016)
F Statistic          358.809*** (df = 9; 1016)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> |
```




```
Linear Regression

1026 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 924, 922, 924, 923, 923, 923, ...
Resampling results:

  RMSE      Rsquared   MAE
  18.45959  0.7582728  13.54923

Tuning parameter 'intercept' was held constant at a value of TRUE
>
```

Net Output Best Model:




```
> NetOutput_regression_fullModel <- lm(noutput ~ size + urea + phosphate + pesticide + pseed + totlabor + wage + goutput + price, data = RiceData)
> coef(NetOutput_regression_fullModel)
  (Intercept)          size          urea     phosphate     pesticide         pseed      totlabor          wage       goutput         price
 27.734486653 -68.859366849   0.108509379   0.365381174   0.005198218  -0.475688169   0.054974527   0.103733828   0.842887088   0.619486123
> |
```

```
Linear Regression

1026 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 923, 923, 924, 923, 923, 924, ...
Resampling results:

  RMSE      Rsquared  MAE
  74.65385  0.99815   45.95576

Tuning parameter 'intercept' was held constant at a value of TRUE
> |
```

5. Prediction and Summary

Price Best Model Prediction:

```
> Price_Prediction_BM <- data.frame(urea = c(800), phosphate = c(500), pesticide = c(200) ,pseed = c(500) , purea= c(200), totlabor = c(1000), wage = c(500), goutput = c(1000), noutput = c(1000))
> predict(price_regression_BM,Price_Prediction_BM)
       1
431.7042
```

```
> predict(price_regression_BM,Price_Prediction_BM)
       1
431.7042
```

Sr No	Predictor	Values
1	Urea	800
2	Phosphate	500
3	Pesticide	200
4	Pseed	500
5	Purea	200
6	Totlabor	1000
7	Wage	500
8	Goutput	1000
9	Noutput	1000
Prediction of Price based on the above values of Predictors		431.7042

- From the above prediction, we can understand that if the predictor values are set according to the above data the price value of Rice will be 431.7042.
- We have summarized the values in a table for a better understanding of our prediction of the Price of Rice.

Net Output Best Model Prediction:

```
Netoutput_Prediction_BM <- data.frame (size = c(150), urea = c(1500), phosphate = c(1000), pesticide = c(60000), pseed = c(500), totlabor = c(1000) , wage= c(300), goutput = c(35000), price = c(400))

predict(NetOutput_regression_BM,Netoutput_Prediction_BM)
```

```
> predict(NetOutput_regression_BM,Netoutput_Prediction_BM)
       1
20135.96
>
```

Sr No	Predictor	Values
1	Size	150
2	Urea	1500
3	Phosphate	1000
4	Pesticide	60000
5	Pseed	500
6	Total labor	1000
7	Wage	300
8	Goutput	1000
9	Price	400
Prediction of Price based on the above values of Predictors		20135.96

- From the above prediction, we can understand that if the predictor values are set according to the above data the net output of production of Rice will be 20135.96.
- We have summarized the values in a table for a better understanding of our prediction of the Net Output of Rice.