The Data Analysis Report about Operating Hours

Cindy 300324658

2023-01-24

Introduction

The Data Analysis Report about Operating Hours The number of working times varies on the work pool and the output level of the various activities. This report takes a large metropolitan department store's worked hours data set as an analysis sample, which initially includes 52 observations and 11 variables, to find a model to help determine the number of working hours required to operate efficiently. For the variables, see the appendix for detail.

Analysis Report

1. Summary Statistics

We are interested in the worked hours and trying to find a suitable model to use explained variables to predict the worked hours. The variable "X", "OBS", and "DAY" of the original data set are not helpful for the analysis. We move them and keep the others. The following table shows the numerical summaries for the response variable worked hours and seven predictor variables.

variables	min	max	median	mean
workedhour	86.6	150.4	117.2	117.4
mail	1832	11777	5542	5586
soldmoney	14	174	90.50	90.98
payment	389	1419	780	782
changeorder	84	577	177	212
cheques	334	1081	546	594
mismail	30	86	57	58
busticket	126	1721	723	754

The table shows the average worked hour for 52 days is about 117 hours. The average number of pieces of mail processed, money orders and gift certificates sod, window payments, change order transactions processed, cheques cashed, pieces of miscellaneous mail processed on an "as available" basis, and bus tickets sold are 5586,91,782,212,594,56,754 respectively.

2. Variables Analysis

The following table shows the correlation between predictors. A number closer to 1 means two variables are more related, and close to 0 are less related. The variables X3, X5, and X7 are moderately related. There are no strongly related variables.

	X1	X2	X3	X4	X5	X6	X7
X1	1.0000000	0.0112820	0.0548036	-0.0431175	-0.2765857	-0.0159404	-0.3117669
X2	0.0112820	1.0000000	0.2452151	0.0368615	-0.0158897	0.3389244	0.1222646
X3	0.0548036	0.2452151	1.0000000	0.4778072	0.5089937	0.3489202	0.5087885
X4	-0.0431175	0.0368615	0.4778072	1.0000000	0.4428052	0.1673518	0.2750750
X5	-0.2765857	-0.0158897	0.5089937	0.4428052	1.0000000	0.3822719	0.5660733
X6	-0.0159404	0.3389244	0.3489202	0.1673518	0.3822719	1.0000000	0.2971547
X7	-0.3117669	0.1222646	0.5087885	0.2750750	0.5660733	0.2971547	1.0000000

The following figure shows that the distribution of response variable Y (work hours) is normal. The distribution of predictors X1, X2, X3, and X6 are normal, and X4, X5 and X7 are skewed to the right. Except for X1 (number of pieces of mail processed), other predictors all positive related to Y. X5(number of cheques cashed), X6(number of pieces of miscellaneous mail processed on an "as available" basis), X7(number of bus tickets sold), and X3(number of window payments) have relatively strong relations with Y. Meanwhile, X3, X4(number of change order transactions processed), X5, and X7 moderately correlated.

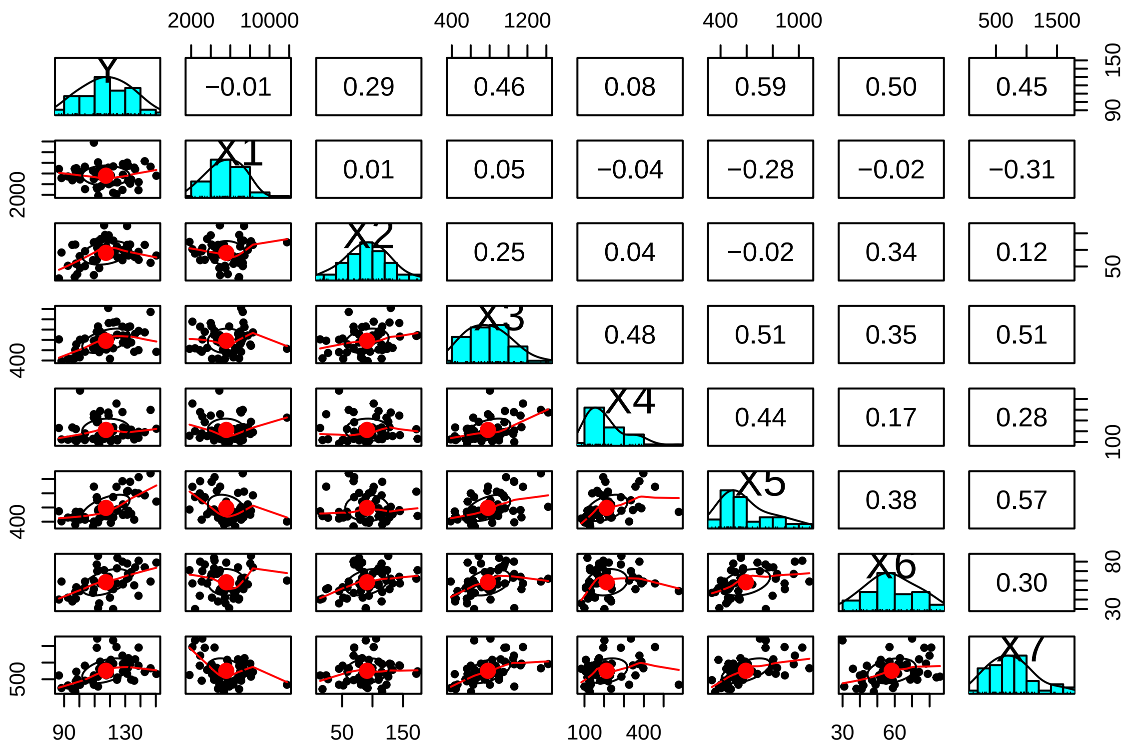


Figure 1: The Correlation of Predicters

3. Modeling

According to the previous figure, There may be some linear relationship between Y and Xs. We begin with an experiment with multiple linear models to fit data. The multiple linear model can be described as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

.

1. Full Linear Regression Model

We construct a linear regression model using all predictors, and the model can be described as

$\hat{Y} = 60.5538 + 0.0014 * X1 + 0.0873 * X2 + 0.0087 * X3 - 0.0428 * X4 + 0.0468 * X5 + 0.2092 * X6 + 0.0048 * X7$

.

(1) model significant check

If the model provides a better fit to the data than a model with no independent variable, the model is significant. We can use F-test to evaluate the overall significance of the model.

-Null hypothesis states that the model with no independent predictors (intercept) fits the data and the full model. -Alternative hypothesis states that the full model fits better than the one with only intercept.

Mathematically:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

H_A : at least one β not equal to 0.

The full model's overall F-statistic value is 8.277, and the p-value is 2.053e-06, much less than 0.001. We can reject the null hypothesis and prefer the alternative hypothesis; at least one β is not equal to 0, and model$\hat{Y}$ perform better than an intercept-only model, which is significant.

(2) predictor significance check

While not all predictors have the same significance. The following table shows the P-value (α) of intercept (β_0) and six predictors.

variable	β_0	X1	X2	X3	X4	X5	X6	X7
p-value	0	0.1481	0.0774	0.3485	0.0176	3e-04	0.1153	0.38664

The p-value of β_0 and predictors X2, X4, and X5 are less than 0.1, which means they are significant at α=0.1 level; the other predictors are insignificant.

The performance of the full model $\hat{Y}$ is not so good, since the R^2_{adj} is about 0.4997 and Residual standard error is about 10.99.

2. Reduced Variables Model

As some variables are not so significant, we select significant variables to build a reduced variables linear regression model. The predictors X2, X4 and X5 are significant at $\alpha = 0.1$. The reduced model based on X2,X4 and X5, can be mathematically stated as :

$$\hat{Y_1} = 77.72564 + 0.13626 * X2 - 0.03469 * X4 + 0.05827 * X5$$

The following table shows the residual standard error (sigma) and adjusted R-square value of the full model and some reduced models. The full model used all seven predictors, "X2+X4+X5," which used predictors X2, X4 and X5, and so on.

criteria	full model	X2+X4+X5	X5	X5+X4	X5+X2	X2+X4
sigma	10.99018	11.54278	12.7007	12.44982	11.90183	15.11134
R^2_{adj}	0.4996966	0.4481208	0.3318425	0.3579782	0.4132525	0.05413445

We can see that the full model performs better than other models, as the sigma is the smallest and the R^2 value is the greatest. The significant predictor model is not as good as the full model but better than other reduced models. For significant model $\hat{Y}_1$, there is no need to take any more predictors.

The full model performs better than the reduced model, but it may contain some correlated and not-so-significant variables. The reduced model has fewer variables; it is easy to focus on significant predictors. While it also likely misses some significant predictors.

As the linear regression models do not work very well, we check the model's assumption.

-the Residuals vs Fitted figure presents a nonlinear effect using $\hat{Y}_1$ reduced model. -the Q-Q plot indicates some departures from normality -the line is cured, and the residual appears to change from left to right. There appears to be some heteroscedasticity. - The Residual vs Leverage figure presents no points outside the dotted line. There are no outliers.

The assumption of linear regression is not satisfied.

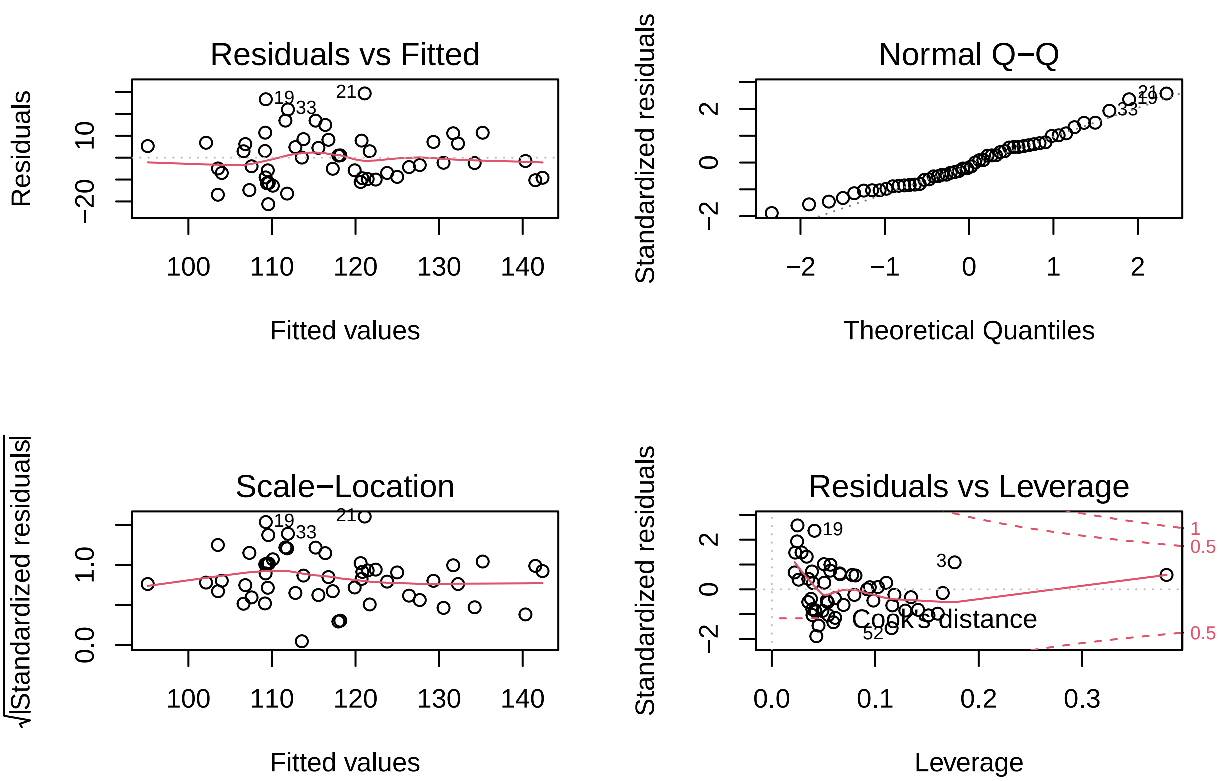


Figure 2: The Plot of Model Y1

3. interaction model

We experiment interaction model. The two-way interaction of significant predictors X2, X4 and X5 can be described as

$$\hat{Y}_2 = 65.97501 - 0.20397 * X2 - 0.02027 * X4 + 0.08380 * X5 + 0.00012 * X2X4 - 0.00017 * X2X5 - 0.00004 * X4X5$$

If we experiment with all predictors and all possible two-way interactions, we will build a model with 28 parameters.

$$\hat{Y}_3 = \beta_0 + \beta 1X_1 + \beta_2X_2 + \beta_3X_3 + ...\beta_7X_7 + \beta_8X1X_2 + \beta_9X1X3......\beta_{28}X_6X_7$$

The X variables' coefficient (Beta) and the P-value(Pr) are shown in the following table. The intercept is β_0. The predictors X2 and X2*X6 are significant in the model, as their p-value is less than 0.1.

	Beta	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.4259675	96.9525338	-0.2003658	0.8429557
X1	0.0021527	0.0096290	0.2235676	0.8250676
X2	1.1220192	0.5949158	1.8860134	0.0719845
X3	0.0244085	0.1045847	0.2333845	0.8175275
X4	-0.0963635	0.2464036	-0.3910801	0.6993373
X5	0.0220598	0.1168031	0.1888635	0.8518564
X6	0.4781270	1.4820118	0.3226202	0.7498974
X7	0.1017618	0.0740021	1.3751203	0.1823422
X1:X2	-0.0000509	0.0000664	-0.7656248	0.4516855
X1:X3	-0.0000024	0.0000108	-0.2212286	0.8268667
X1:X4	-0.0000006	0.0000234	-0.0265315	0.9790623
X1:X5	-0.0000051	0.0000109	-0.4683584	0.6439366
X1:X6	0.0001789	0.0001361	1.3141872	0.2017397
X1:X7	-0.0000019	0.0000057	-0.3290931	0.7450633
X2:X3	0.0002738	0.0006830	0.4008686	0.6922165
X2:X4	-0.0003436	0.0011704	-0.2935442	0.7717387
X2:X5	0.0000372	0.0005719	0.0651065	0.9486518
X2:X6	-0.0174120	0.0072599	-2.3983887	0.0249722
X2:X7	0.0000911	0.0003931	0.2316650	0.8188469
X3:X4	0.0001646	0.0001693	0.9720478	0.3411393
X3:X5	0.0000032	0.0001283	0.0250757	0.9802109
X3:X6	-0.0004021	0.0009675	-0.4156691	0.6815048
X3:X7	-0.0000663	0.0000503	-1.3177487	0.2005634
X4:X5	0.0000091	0.0003306	0.0276266	0.9781983
X4:X6	-0.0005886	0.0029076	-0.2024190	0.8413691
X4:X7	-0.0000153	0.0001107	-0.1385530	0.8910094
X5:X6	0.0016350	0.0018265	0.8951449	0.3799824
X5:X7	-0.0000692	0.0000765	-0.9047218	0.3749925
X6:X7	-0.0000670	0.0005700	-0.1175273	0.9074625

According to the previous result, we try to build models with X2, X3, X4, X5, X6 and X2X6. The model can be described as

$$\hat{Y}_4 = 39.7754 + 0.4379 * X2 + 0.0131 * X3 - 0.0427 * X4 + 0.0423 * X5 + 0.7898 * X6 - 0.0063X2 * X6$$

.

We compare the performer for the four models and find out the modelY_4 perform the best, as its RSE is the smallest and adjusted R square value is the greatest. So, we choose model Y_4 to study.

Model	RSE	R_square	Adjusted_R	Note
Y	10.99018	0.5683657	0.4996966	lm(formula = Y ~ ., data = cler)
Y1	11.54278	0.4805843	0.4481208	lm(formula = Y ~ X2 + X4 + X5, data = cler)
Y2	11.87068	0.4849882	0.4163200	lm(formula = Y ~ X2 + X4 + X5 + X2 * X4 + X2 * X5 + X4 * X5, data = cler)
Y3	11.35785	0.7590241	0.4656621	lm(formula = Y ~ .^2, data = cler)
Y4	10.84006	0.5705338	0.5132717	lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X2 * X6, data = cler)

There is a two-way interaction among the predictors in model Y_4,X2X6.To reference whether the predictor X2*X6 is significant,we hypotheses,

$H_0 : \beta_6 = 0$ The coefficient of the X2X6 is zero.

$H_A : \beta_6 \neq 0$ The coefficient of the X2X6 is not zero.

The t value of X2:X6 is -1.636, the p-value is about 0.1087 that greater than 0.1. It is not significant at $\alpha = 0.1$ level.

We reduce the variable X2X6 and make a new model. The model can be described as

$$\hat{Y_5} = 68.27443 + 0.08309 * X2 + 0.01386 * X3 - 0.04345 * X4 + 0.04471 * X5 + 0.22909 * X6$$

.

The following table shows the performance of model Y4 and reduced interaction model Y5. We can see the R^2_{adj} value of the Y5 is 0.4955, smaller than the Y4; when we reduced the interaction variable, the performance decreased. The performance of model Y5 is worse than Y4.

Model	RSE	R_square	Adjusted_R	Note
Y4	10.84	0.571	0.5133	lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X2 * X6, data = cler)
Y5	11.04	0.545	0.4955	lm(formula = Y ~ X2 + X3 + X4 + X5 + X6, data = cler)

Conclusion

In conclusion, the model Y_4 performs the best. The second better model is the full model Y. However, the models' performance needs to be better. It is better to deal with the problem of not suiting the assumptions of linear regression or try another machine learning engine to get better predictive power.

Appendix

step 1: deal with sample data set

```
library(tidyverse)
library(dplyr)

clerical<-read.csv("clerical.csv")
cler<-clerical[,-(1:3)]

cler1<-rename(cler,workedhour = Y,mail=X1,soldmoney=X2,payment=X3,changeorder=X4,cheques=X5,mismail=X6,
str(cler)     # rename the variables
```

```
## 'data.frame':    52 obs. of  8 variables:
##  $ Y : num  128 114 147 124 100 ...
##  $ X1: int  7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
##  $ X2: int  100 110 61 102 45 144 123 78 172 126 ...
##  $ X3: int  886 962 1342 1153 803 1127 627 748 876 685 ...
##  $ X4: int  235 388 398 457 577 345 326 161 219 287 ...
##  $ X5: int  644 589 1081 891 537 563 402 495 823 555 ...
##  $ X6: int  56 57 59 57 49 64 60 57 62 86 ...
##  $ X7: int  737 1029 830 1468 335 918 335 962 665 577 ...
```

step 2: summary the data set

```
summary(cler1)
```

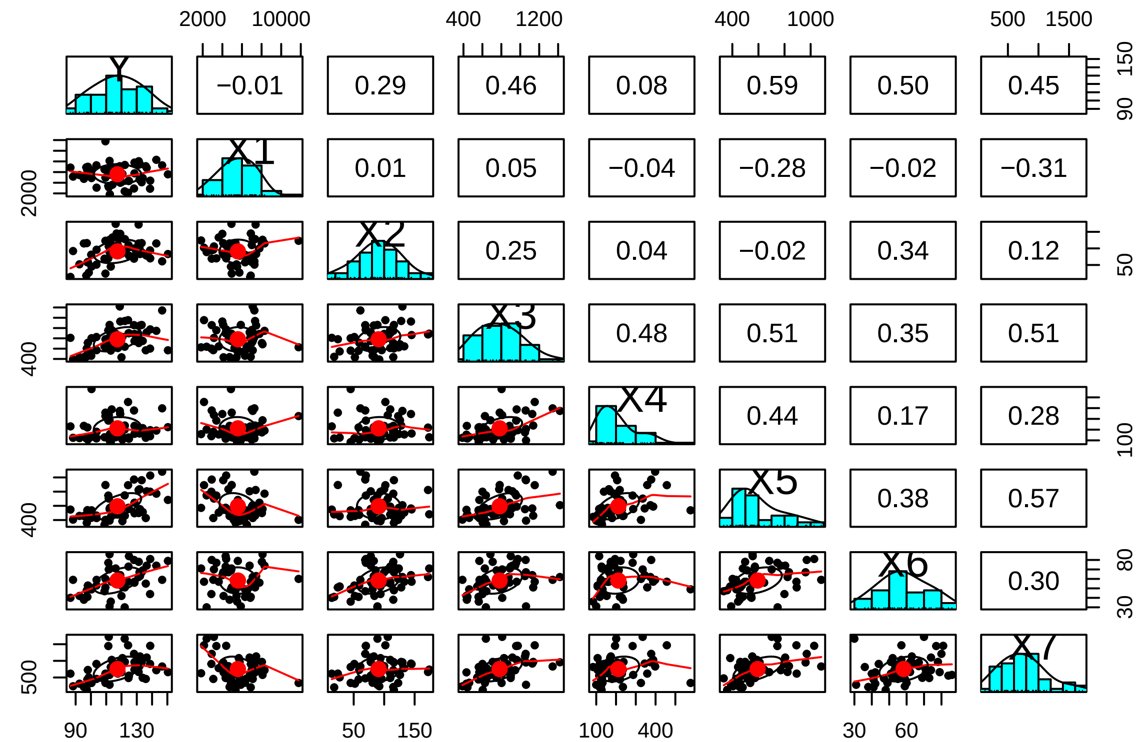
```
##    workedhour         mail         soldmoney         payment
##  Min.   : 86.6   Min.   : 1832   Min.   : 14.00   Min.   : 389.0
##  1st Qu.:107.6   1st Qu.: 4338   1st Qu.: 69.75   1st Qu.: 603.2
##  Median :117.2   Median : 5542   Median : 90.50   Median : 780.0
##  Mean   :117.4   Mean   : 5586   Mean   : 90.98   Mean   : 782.3
##  3rd Qu.:129.2   3rd Qu.: 6996   3rd Qu.:115.00   3rd Qu.: 924.2
##  Max.   :150.4   Max.   :11777   Max.   :174.00   Max.   :1419.0
##   changeorder       cheques          mismail        busticket
##  Min.   : 84.0   Min.   : 334.0   Min.   :30.00   Min.   : 126.0
##  1st Qu.:128.5   1st Qu.: 460.2   1st Qu.:49.00   1st Qu.: 481.2
##  Median :177.0   Median : 545.5   Median :57.00   Median : 722.5
##  Mean   :211.9   Mean   : 593.9   Mean   :58.27   Mean   : 753.8
##  3rd Qu.:235.8   3rd Qu.: 712.8   3rd Qu.:69.25   3rd Qu.: 958.2
##  Max.   :577.0   Max.   :1081.0   Max.   :86.00   Max.   :1721.0
```

```
cor<-cor(cler[,-1])

knitr::kable(cor,full_width=FALSE)
```

	X1	X2	X3	X4	X5	X6	X7
X1	1.0000000	0.0112820	0.0548036	-0.0431175	-0.2765857	-0.0159404	-0.3117669
X2	0.0112820	1.0000000	0.2452151	0.0368615	-0.0158897	0.3389244	0.1222646
X3	0.0548036	0.2452151	1.0000000	0.4778072	0.5089937	0.3489202	0.5087885

	X1	X2	X3	X4	X5	X6	X7
X4	-0.0431175	0.0368615	0.4778072	1.0000000	0.4428052	0.1673518	0.2750750
X5	-0.2765857	-0.0158897	0.5089937	0.4428052	1.0000000	0.3822719	0.5660733
X6	-0.0159404	0.3389244	0.3489202	0.1673518	0.3822719	1.0000000	0.2971547
X7	-0.3117669	0.1222646	0.5087885	0.2750750	0.5660733	0.2971547	1.0000000

```
library(psych)

pairs.panels(cler, method = "pearson")
```



growth_0<-growth %>% pivot_longer(c(“PCE”,“GCE”,“EXP”,“IMP”,“POPULATION”,“MBASE”,“PI”,“UNEMR”,“FDIV =“levels”,values_to = “values”) growth_0 %>% ggplot(aes(DATE,values))+ geom_line(color=“chocolate”)+ facet_wrap(~levels,scales = “free_y”,ncol = 3)+ ggtitle(“Relavant Ecomical Factors Change from 1999 to 2021”)

step 3: Modeling

```
# full model

lm_full<-lm(Y~.,cler)
summary(lm_full)
```

```
##
## Call:
## lm(formula = Y ~ ., data = cler)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.537  -7.038  -1.224   6.168  28.012
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.5537920  9.4952130   6.377  9.4e-08 ***
## X1           0.0013496  0.0009168   1.472  0.14813
## X2           0.0872715  0.0482561   1.809  0.07736 .
## X3           0.0086879  0.0091681   0.948  0.34850
## X4          -0.0427781  0.0173449  -2.466  0.01762 *
## X5           0.0467902  0.0119808   3.905  0.00032 ***
## X6           0.2092130  0.1302236   1.607  0.11530
## X7           0.0048192  0.0055105   0.875  0.38657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.99 on 44 degrees of freedom
## Multiple R-squared:  0.5684, Adjusted R-squared:  0.4997
## F-statistic: 8.277 on 7 and 44 DF,  p-value: 2.053e-06
```

```r
# f-statistic

summary(lm_full)$fstatistic
```

```
##     value     numdf     dendf
##  8.276877  7.000000 44.000000
```

```r
# p-value of the predictors
p<-round(as.data.frame(summary(lm_full)$coefficients)[,4],4)
a<-c("intercept","X1","X2","X3","X4","X5","X6","X7")
rbind(a,p)
```

```
##   [,1]        [,2]     [,3]     [,4]     [,5]     [,6]    [,7]     [,8]
## a "intercept" "X1"     "X2"     "X3"     "X4"     "X5"    "X6"     "X7"
## p "0"         "0.1481" "0.0774" "0.3485" "0.0176" "3e-04" "0.1153" "0.3866"
```

```r
# evaluate the model

summary(lm_full)$r.sq
```

```
## [1] 0.5683657
```

```r
summary(lm_full)$sigma
```

```
## [1] 10.99018
```

```r
sqrt(summary(lm_full)$sigma)
```

```
## [1] 3.315145
```

```r
# reduced model

lm_1<- lm(Y~X2+X4+X5,cler)
lm_1
```

```
##
## Call:
## lm(formula = Y ~ X2 + X4 + X5, data = cler)
##
## Coefficients:
## (Intercept)           X2           X4           X5
##    77.72564      0.13626     -0.03469      0.05827
```

```r
summary(lm_1)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X4 + X5, data = cler)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.259  -9.075  -1.938   6.882  29.303
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 77.725640   6.910199  11.248 4.69e-15 ***
## X2           0.136264   0.045413   3.001  0.00426 **
## X4          -0.034689   0.017140  -2.024  0.04857 *
## X5           0.058268   0.009714   5.998 2.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.54 on 48 degrees of freedom
## Multiple R-squared:  0.4806, Adjusted R-squared:  0.4481
## F-statistic:  14.8 on 3 and 48 DF,  p-value: 5.91e-07
```

```r
lm_2<-lm(Y~X5,cler)
lm_3<-lm(Y~X5+X4,cler)
lm_4<-lm(Y~X5+X2,cler)
lm_5<-lm(Y~X2+X4,cler)



lm_6<-lm(Y~X2+X4+X5+X2*X4+X2*X5+X4*X5, cler)
round(lm_6$coefficients,5)
```

```
## (Intercept)          X2          X4          X5       X2:X4       X2:X5
##    65.97501     0.20397    -0.02027     0.08380     0.00012    -0.00017
##       X4:X5
##    -0.00004
```

```r
lm_7<-lm(Y~.^2,cler)
summary(lm_7)
```

```
##
## Call:
## lm(formula = Y ~ .^2, data = cler)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.6874  -4.6622  -0.3905   4.4598  18.4249
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.943e+01  9.695e+01  -0.200    0.843
## X1           2.153e-03  9.629e-03   0.224    0.825
## X2           1.122e+00  5.949e-01   1.886    0.072 .
## X3           2.441e-02  1.046e-01   0.233    0.818
## X4          -9.636e-02  2.464e-01  -0.391    0.699
## X5           2.206e-02  1.168e-01   0.189    0.852
## X6           4.781e-01  1.482e+00   0.323    0.750
## X7           1.018e-01  7.400e-02   1.375    0.182
## X1:X2       -5.087e-05  6.645e-05  -0.766    0.452
## X1:X3       -2.382e-06  1.077e-05  -0.221    0.827
## X1:X4       -6.219e-07  2.344e-05  -0.027    0.979
## X1:X5       -5.106e-06  1.090e-05  -0.468    0.644
## X1:X6        1.789e-04  1.361e-04   1.314    0.202
## X1:X7       -1.864e-06  5.663e-06  -0.329    0.745
## X2:X3        2.738e-04  6.830e-04   0.401    0.692
## X2:X4       -3.436e-04  1.170e-03  -0.294    0.772
## X2:X5        3.724e-05  5.719e-04   0.065    0.949
## X2:X6       -1.741e-02  7.260e-03  -2.398    0.025 *
## X2:X7        9.107e-05  3.931e-04   0.232    0.819
## X3:X4        1.646e-04  1.693e-04   0.972    0.341
## X3:X5        3.217e-06  1.283e-04   0.025    0.980
## X3:X6       -4.021e-04  9.675e-04  -0.416    0.682
## X3:X7       -6.634e-05  5.034e-05  -1.318    0.201
## X4:X5        9.132e-06  3.306e-04   0.028    0.978
## X4:X6       -5.886e-04  2.908e-03  -0.202    0.841
## X4:X7       -1.534e-05  1.107e-04  -0.139    0.891
## X5:X6        1.635e-03  1.826e-03   0.895    0.380
## X5:X7       -6.920e-05  7.649e-05  -0.905    0.375
## X6:X7       -6.699e-05  5.700e-04  -0.118    0.907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.36 on 23 degrees of freedom
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.4657
## F-statistic: 2.587 on 28 and 23 DF,  p-value: 0.0114
```

```r
data_1<-as.data.frame(summary(lm_7)$coefficient)
data_1<-rename(data_1,Beta=Estimate)

knitr::kable(data_1,full_width=FALSE)
```

	Beta	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.4259675	96.9525338	-0.2003658	0.8429557
X1	0.0021527	0.0096290	0.2235676	0.8250676
X2	1.1220192	0.5949158	1.8860134	0.0719845
X3	0.0244085	0.1045847	0.2333845	0.8175275
X4	-0.0963635	0.2464036	-0.3910801	0.6993373
X5	0.0220598	0.1168031	0.1888635	0.8518564
X6	0.4781270	1.4820118	0.3226202	0.7498974
X7	0.1017618	0.0740021	1.3751203	0.1823422
X1:X2	-0.0000509	0.0000664	-0.7656248	0.4516855
X1:X3	-0.0000024	0.0000108	-0.2212286	0.8268667
X1:X4	-0.0000006	0.0000234	-0.0265315	0.9790623
X1:X5	-0.0000051	0.0000109	-0.4683584	0.6439366
X1:X6	0.0001789	0.0001361	1.3141872	0.2017397
X1:X7	-0.0000019	0.0000057	-0.3290931	0.7450633
X2:X3	0.0002738	0.0006830	0.4008686	0.6922165
X2:X4	-0.0003436	0.0011704	-0.2935442	0.7717387
X2:X5	0.0000372	0.0005719	0.0651065	0.9486518
X2:X6	-0.0174120	0.0072599	-2.3983887	0.0249722
X2:X7	0.0000911	0.0003931	0.2316650	0.8188469
X3:X4	0.0001646	0.0001693	0.9720478	0.3411393
X3:X5	0.0000032	0.0001283	0.0250757	0.9802109
X3:X6	-0.0004021	0.0009675	-0.4156691	0.6815048
X3:X7	-0.0000663	0.0000503	-1.3177487	0.2005634
X4:X5	0.0000091	0.0003306	0.0276266	0.9781983
X4:X6	-0.0005886	0.0029076	-0.2024190	0.8413691
X4:X7	-0.0000153	0.0001107	-0.1385530	0.8910094
X5:X6	0.0016350	0.0018265	0.8951449	0.3799824
X5:X7	-0.0000692	0.0000765	-0.9047218	0.3749925
X6:X7	-0.0000670	0.0005700	-0.1175273	0.9074625

```
lm_8<-lm(Y~X2+X3+X4+X5+X6+X2*X6,cler)
summary(lm_7)
```

```
##
## Call:
## lm(formula = Y ~ .^2, data = cler)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.6874  -4.6622  -0.3905   4.4598  18.4249
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.943e+01  9.695e+01  -0.200    0.843
## X1           2.153e-03  9.629e-03   0.224    0.825
## X2           1.122e+00  5.949e-01   1.886    0.072 .
## X3           2.441e-02  1.046e-01   0.233    0.818
## X4          -9.636e-02  2.464e-01  -0.391    0.699
## X5           2.206e-02  1.168e-01   0.189    0.852
## X6           4.781e-01  1.482e+00   0.323    0.750
## X7           1.018e-01  7.400e-02   1.375    0.182
```

```
## X1:X2       -5.087e-05  6.645e-05  -0.766    0.452
## X1:X3       -2.382e-06  1.077e-05  -0.221    0.827
## X1:X4       -6.219e-07  2.344e-05  -0.027    0.979
## X1:X5       -5.106e-06  1.090e-05  -0.468    0.644
## X1:X6        1.789e-04  1.361e-04   1.314    0.202
## X1:X7       -1.864e-06  5.663e-06  -0.329    0.745
## X2:X3        2.738e-04  6.830e-04   0.401    0.692
## X2:X4       -3.436e-04  1.170e-03  -0.294    0.772
## X2:X5        3.724e-05  5.719e-04   0.065    0.949
## X2:X6       -1.741e-02  7.260e-03  -2.398    0.025 *
## X2:X7        9.107e-05  3.931e-04   0.232    0.819
## X3:X4        1.646e-04  1.693e-04   0.972    0.341
## X3:X5        3.217e-06  1.283e-04   0.025    0.980
## X3:X6       -4.021e-04  9.675e-04  -0.416    0.682
## X3:X7       -6.634e-05  5.034e-05  -1.318    0.201
## X4:X5        9.132e-06  3.306e-04   0.028    0.978
## X4:X6       -5.886e-04  2.908e-03  -0.202    0.841
## X4:X7       -1.534e-05  1.107e-04  -0.139    0.891
## X5:X6        1.635e-03  1.826e-03   0.895    0.380
## X5:X7       -6.920e-05  7.649e-05  -0.905    0.375
## X6:X7       -6.699e-05  5.700e-04  -0.118    0.907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.36 on 23 degrees of freedom
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.4657
## F-statistic: 2.587 on 28 and 23 DF,  p-value: 0.0114
```

```
summary(lm_8)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X2 * X6, data = cler)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.957  -6.952  -1.226   7.069  26.559
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.775447  18.987308   2.095 0.041845 *
## X2           0.437936   0.221954   1.973 0.054646 .
## X3           0.013111   0.008290   1.582 0.120766
## X4          -0.042723   0.017070  -2.503 0.016020 *
## X5           0.042278   0.010619   3.981 0.000247 ***
## X6           0.789802   0.365697   2.160 0.036161 *
## X2:X6       -0.006325   0.003865  -1.636 0.108721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.84 on 45 degrees of freedom
## Multiple R-squared:  0.5705, Adjusted R-squared:  0.5133
## F-statistic: 9.964 on 6 and 45 DF,  p-value: 5.501e-07
```

```r
summary(lm_full)
```

```
##
## Call:
## lm(formula = Y ~ ., data = cler)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.537  -7.038  -1.224   6.168  28.012
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.5537920  9.4952130   6.377  9.4e-08 ***
## X1           0.0013496  0.0009168   1.472  0.14813
## X2           0.0872715  0.0482561   1.809  0.07736 .
## X3           0.0086879  0.0091681   0.948  0.34850
## X4          -0.0427781  0.0173449  -2.466  0.01762 *
## X5           0.0467902  0.0119808   3.905  0.00032 ***
## X6           0.2092130  0.1302236   1.607  0.11530
## X7           0.0048192  0.0055105   0.875  0.38657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.99 on 44 degrees of freedom
## Multiple R-squared:  0.5684, Adjusted R-squared:  0.4997
## F-statistic: 8.277 on 7 and 44 DF,  p-value: 2.053e-06
```

```r
model_list<-list(lm_full,lm_1,lm_7,lm_8)
rse<-sapply(model_list,sigma)  #get sigma
       # get r.squre

sum_1<-sapply(model_list,summary)

r_square<-unlist((sum_1[8,]))
r_adj<-unlist((sum_1[9,]))
form<-paste(sum_1[1,])


df<-data.frame(Model=c("Y1","Y2","Y3","Y4"),
               RSE=rse,
               R_square=r_square,
               Adjusted_R=r_adj,
               Note=form)

knitr::kable(df,full_width=FALSE)
```

Model	RSE	R_square	Adjusted_R	Note
Y1	10.99018	0.5683657	0.4996966	lm(formula = Y ~ ., data = cler)
Y2	11.54278	0.4805843	0.4481208	lm(formula = Y ~ X2 + X4 + X5, data = cler)
Y3	11.35785	0.7590241	0.4656621	lm(formula = Y ~ .^2, data = cler)

Model	RSE	R_square	Adjusted_R	Note
Y4	10.84006	0.5705338	0.5132717	lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X2 * X6, data = cler)

```r
lm_9<-lm(Y~X2+X3+X4+X5+X6,cler)
summary(lm_9)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3 + X4 + X5 + X6, data = cler)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.249  -7.439  -1.807   7.619  28.406
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.274431   7.701975   8.865 1.63e-11 ***
## X2           0.083086   0.048224   1.723  0.09162 .
## X3           0.013864   0.008427   1.645  0.10674
## X4          -0.043445   0.017372  -2.501  0.01602 *
## X5           0.044711   0.010705   4.177  0.00013 ***
## X6           0.229095   0.130120   1.761  0.08495 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.04 on 46 degrees of freedom
## Multiple R-squared:  0.545,  Adjusted R-squared:  0.4955
## F-statistic: 11.02 on 5 and 46 DF,  p-value: 5.19e-07
```

```r
dataf<-data.frame(Model=c("Y4","Y5"),
               RSE=c("10.84","11.04"),
               R_square=c("0.571","0.545"),
               Adjusted_R=c("0.5133","0.4955"),
               Note=c("lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X2 * X6, data = cler)",
                      "lm(formula = Y ~ X2 + X3 + X4 + X5 + X6, data = cler)"))

knitr::kable(dataf,full_width=FALSE)
```

Model	RSE	R_square	Adjusted_R	Note
Y4	10.84	0.571	0.5133	lm(formula = Y ~ X2 + X3 + X4 + X5 + X6 + X2 * X6, data = cler)
Y5	11.04	0.545	0.4955	lm(formula = Y ~ X2 + X3 + X4 + X5 + X6, data = cler)