# Supervisor Performance
# Fall 2020

MSIS 545 711 20F
Lokesh Kumar Naragani
Shashank Kala

Prof. J. E. Helmreich

## 1. Introduction

We were given the data which involves the analysis of a survey of clerical employees in a large financial organization. We wish to explore the relationship between overall satisfaction with a specific supervisor and answers given to six selected questions on the survey. This dataset consists of the percentage of the positive responses in each of the 30 departments headed by a specific supervisor. Each department had about 35 employees responded with a rating in the range 1-5 for all the 7 questions(variables) and the responses of (1,2) are considered in favor of the supervisor performance.

## 2. Getting the Data into R

Firstly, loading the dataset (sup.R) file into the RStudio and the choose the file downloaded

```
> sup <- read.table(file.choose())
```

We see that sup dataset is loaded into the RStudio with the 30 observations (i.e., 30 departments) of 7 variables (i.e., 7 survey questions asked to 35 employees in each department).

## 3. Descriptive Analysis

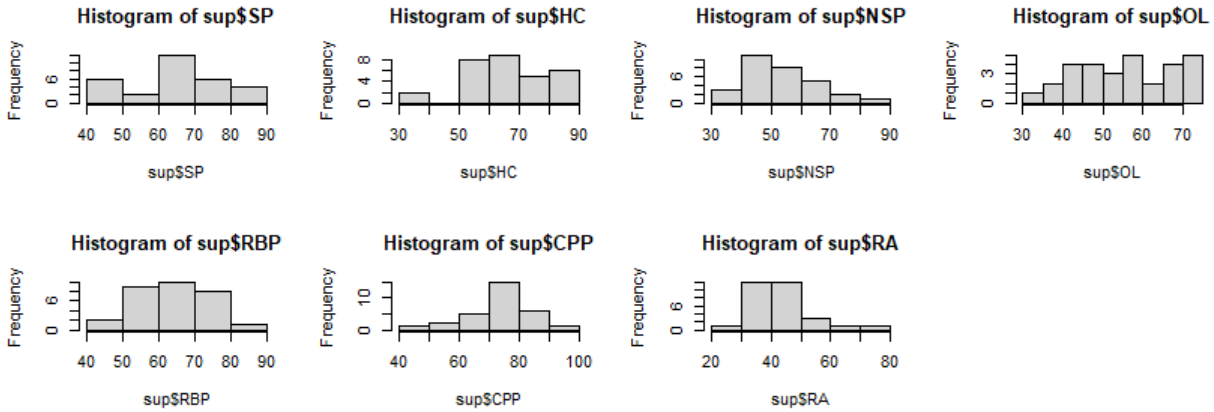
Let us now at the summary statistics of the data to get better understanding of the dataset and how it is spread.

```
> summary(sup)
 SP              HC              NSP             OL
 Min.   :40.00   Min.   :37.0   Min.   :30.00   Min.  :34.00
 1st Qu.:58.75   1st Qu.:58.5   1st Qu.:45.00   1stQu.:47.00
 Median :65.50   Median :65.0   Median :51.50   Median:56.50
 Mean   :64.63   Mean   :66.6   Mean   :53.13   Mean  :56.37
 3rd Qu.:71.75   3rd Qu.:77.0   3rd Qu.:62.50   3rdQu.:66.75
 Max.   :85.00   Max.   :90.0   Max.   :83.00   Max.  :75.00
                    RBP             CPP             RA
 Min.   :43.00   Min.   :49.00   Min.   :25.00
 1st Qu.:58.25   1st Qu.:69.25   1st Qu.:35.00
 Median :63.50   Median :77.50   Median :41.00
Mean   :64.63   Mean   :74.77   Mean   :42.93
 3rd Qu.:71.00   3rd Qu.:80.00   3rd Qu.:47.75
 Max.   :88.00   Max.   :92.00   Max.   :72.00
```

Let us create the histogram and scatterplots for visualization:

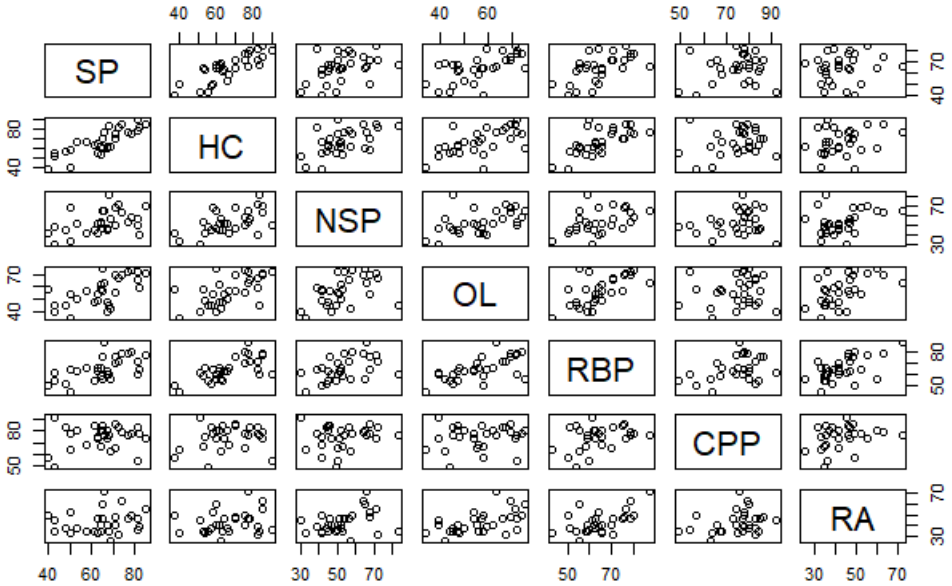
```
> hist(sup$SP)
> hist(sup$HC)
> hist(sup$NSP)
```

```
> hist(sup$OL)
> hist(sup$RBP)
> hist(sup$CPP)
> hist(sup$RA)
```




Scatterplot:

```
> plot(sup,1)
```




**Inferences:**

1, For overall job rating or supervisor performance (SP) 12 departments had 60-70 percentage of positive responses, and only 4 departments had 80 percentage of positive responses.
2.80 plus Percentage of employees in 6 departments said that complaint was properly handled by the supervisor.
3.Only in 1 department 80 percentage plus of employees said that No special privileges were given.
4.In 5 department 70 percent and above employees agreed that there were opportunities to learn new things.
5.In 9 departments 70 percent and above employees agreed that there was salary raise based on performance.
6.In 21 departments 70 percent plus employees said that supervisor is too critical of poor performance.
7.Only 1 department had majority of employees claiming it is advancement to better jobs.

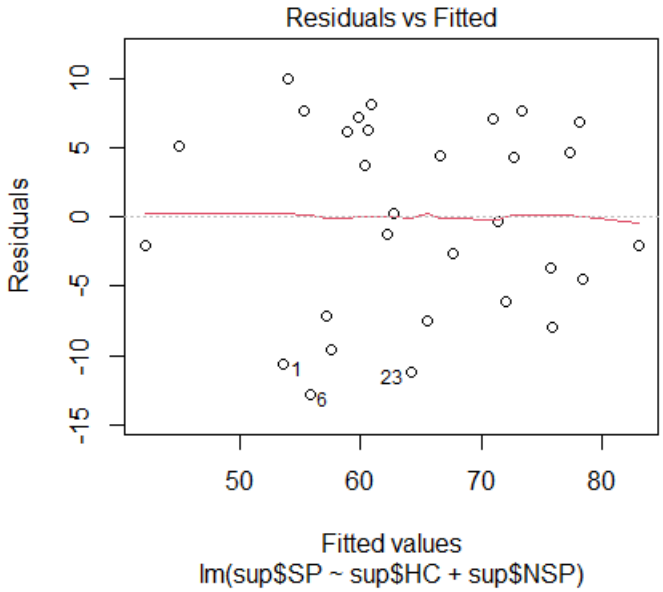
## 4. Coefficient Interpretation

Let us fit the linear model of SP against HC and NSP as shown below:

```
> suplm <- lm(sup$SP ~ sup$HC + sup$NSP)
> summary(suplm)

Call:
lm(formula = sup$SP ~ sup$HC + sup$NSP)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7887  -5.6893  -0.0284   6.2745   9.9726

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.32762    7.16023   2.141   0.0415 *
sup$HC       0.78034    0.11939   6.536 5.22e-07 ***
sup$NSP     -0.05016    0.12992  -0.386   0.7025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.102 on 27 degrees of freedom
Multiple R-squared:  0.6831,  Adjusted R-squared:  0.6596
F-statistic:  29.1 on 2 and 27 DF,  p-value: 1.833e-07
```

Residual Plot of the linear model:

```
> plot(suplm,1)
```




**Inference:**

1.The coefficient for HC is 0.78 whereas for NSP is -.05 , hence while calculating overall supervisor performance we can say th

2. The p-value is 1.833e-07 which is very low hence we can say that the model is significant and changes in the value of HC and NSP are related to change in the Supervisor performance.

3. The residuals bounce randomly around the residual = 0 line. This is suggesting that the relationship is linear.

## 5.Full model

Let us create the model using all six covariates.

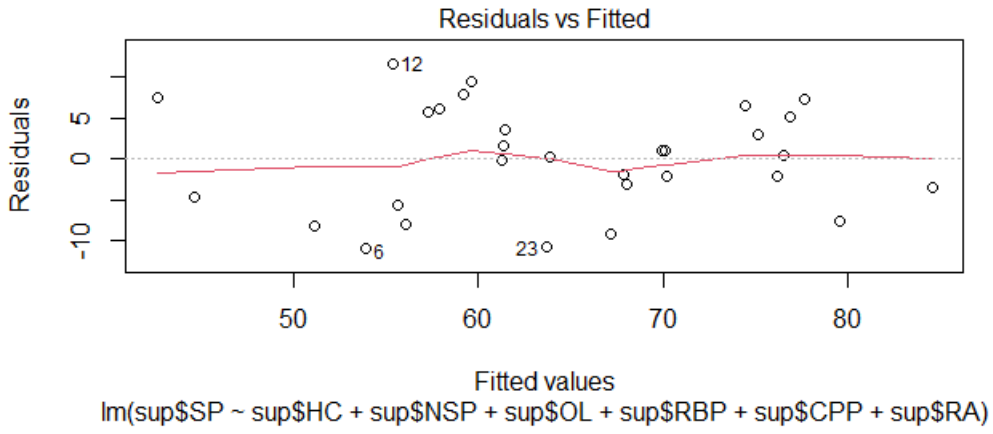
```
> suplm1 <- lm(sup$SP ~ sup$HC + sup$NSP + sup$OL + sup$RBP +
sup$CPP + sup$RA)
> summary(suplm1)

Call:
lm(formula = sup$SP ~ sup$HC + sup$NSP + sup$OL + sup$RBP +
sup$CPP +
    sup$RA)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
sup$HC       0.61319    0.16098   3.809 0.000903 ***
sup$NSP     -0.07305    0.13572  -0.538 0.595594
sup$OL       0.32033    0.16852   1.901 0.069925 .
sup$RBP      0.08173    0.22148   0.369 0.715480
sup$CPP      0.03838    0.14700   0.261 0.796334
sup$RA      -0.21706    0.17821  -1.218 0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,  Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

The p-value we get here is exceedingly small which proves that the significant difference does exist between the mean of variables HC,NSP,OL,RBP,CPP,RA.

Let us plot the residuals:

```
> plot(suplm1,1)
```




**Confidence interval:**

Let us calculate the confidence interval

```
> confint(suplm1)
```

```
                   2.5 %     97.5 %
(Intercept) -13.18712881 34.7612816
sup$HC        0.28016866  0.9462066
sup$NSP      -0.35381806  0.2077178
sup$OL       -0.02827872  0.6689430
sup$RBP      -0.37642935  0.5398936
sup$CPP      -0.26570179  0.3424647
sup$RA       -0.58571106  0.1515977
```

**Prediction Interval:**

We calculate the prediction interval for Supervisor Performance (SP) when all the covariates are at the mean level.

```
> newdata = data.frame(HC=66.6, NSP=53.13, OL=56.37, RBP=64.63,
CPP=74.77, RA=42.93)
> predict(suplm1, newdata = newdata, interval = "confidence")
        fit      lwr      upr
1  51.11030 42.55502 59.66557
2  61.35277 57.97029 64.73524
3  69.93944 63.44979 76.42909
4  61.22684 55.28897 67.16471
5  74.45380 70.02428 78.88332
6  53.94185 45.82386 62.05984
7  67.14841 61.99316 72.30367
8  70.09701 64.79421 75.39980
9  79.53099 71.22222 87.83975
10 59.19846 55.18008 63.21683
11 57.92572 51.63028 64.22116
12 55.40103 49.94035 60.86171
13 59.58168 52.64531 66.51805
14 70.21401 59.46431 80.96372
15 76.54933 70.68118 82.41748
16 84.54785 73.82102 95.27468
17 76.15013 69.29093 83.00933
18 61.39736 50.62090 72.17383
19 68.01656 62.66507 73.36805
20 55.62014 49.16527 62.07502
21 42.60324 35.50593 49.70055
22 63.81902 58.92819 68.70985
23 63.66400 59.88479 67.44321
24 44.62475 35.24662 54.00288
25 57.31710 50.76233 63.87187
26 67.84347 57.59134 78.09561
27 75.14036 69.09798 81.18275
```

```
28 56.04535 49.90540 62.18531
29 77.66053 71.98300 83.33806
30 76.87850 69.00992 84.74707
```

We can see the prediction intervals of each of the 30 departments as shown.

**Inference:**

1.The coefficient for HC, OL, RBP,CPP are positive whereas for NSP and RA it is negative , hence while calculating overall supervisor performance we can say that HC(Handled employee Complaints), OL(Opportunity to Learn new things), RBP (Raises Based on Performance), CPP(Critical of Poor Performance) are directly related and NSP(No Special Privileges) , RA(Rate of Advancement) are inversely related.

2. The p-value is 1.24e-05 which is very low hence we can say that the significant differences do exist between the variables while predicting the Supervisor performance.

3. The residuals show there is a linear relationship although when we have more predictor variables, the model shows less linearity.

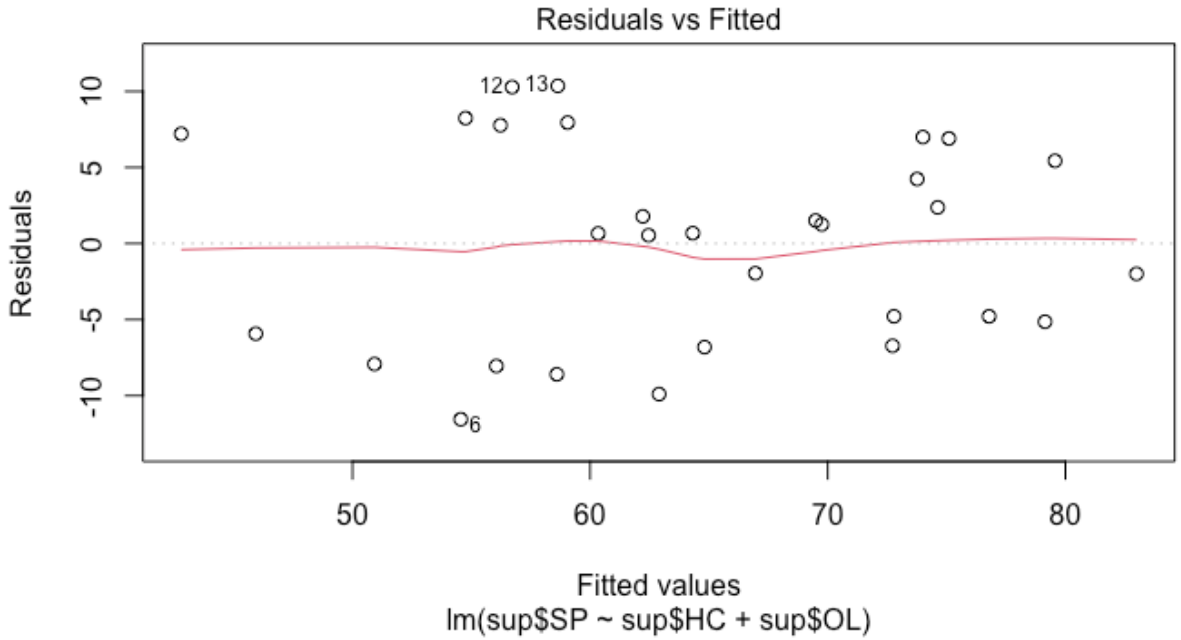
## 6.Reduced Model one

Let us create first reduced Model one using SP on HC and OL as show below:

```
> suplm2 <- lm(sup$SP ~ sup$HC + sup$OL)
> summary(suplm2)
Call:
lm(formula = sup$SP ~ sup$HC + sup$OL)

Residuals:
    Min      1Q  Median      3Q     Max
-11.804  -5.831   1.374   4.328  10.625

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.203424  31.733940   0.983    0.335
sup$HC         0.306661   0.502732   0.610    0.547
sup$OL        -0.170646   0.569888  -0.299    0.767
sup$HC:sup$OL  0.005886   0.008531   0.690    0.496

Residual standard error: 6.884 on 26 degrees of freedom
Multiple R-squared:  0.7133,    Adjusted R-squared:  0.6802
F-statistic: 21.56 on 3 and 26 DF,  p-value: 3.174e-07
```

The P-value is very small showing that there is the difference between the means of HC and OL

```
> plot(suplm2,1)
```




```
> confint(suplm2)
                  2.5 %     97.5 %
(Intercept) -4.61755361 24.3593145
sup$HC       0.40042203  0.8866132
sup$OL      -0.06458185  0.4869655
> newdata1 = data.frame(HC=66.6,OL=56.37)
> predict(suplm2, newdata1, interval="confidence")
        fit      lwr      upr
1  50.92676 46.21895 55.63457
2  62.46037 59.84231 65.07842
3  69.48935 65.50119 73.47751
4  60.33851 56.98271 63.69430
5  74.00392 70.47233 77.53550
6  54.55679 50.73237 58.38121
7  64.81330 62.25352 67.37309
8  69.75025 66.32222 73.17829
9  76.78918 72.78502 80.79333
10 59.05147 55.75756 62.34538
11 56.22644 51.81832 60.63456
12 56.71842 51.93703 61.49981
13 58.63903 54.37511 62.90295
14 72.78648 65.91944 79.65351
15 74.62755 70.32739 78.92770
16 82.99328 77.68317 88.30339
17 79.14211 74.66240 83.62182
```

```
18 64.32132 57.58783 71.05481
19 66.95505 64.29752 69.61258
20 58.59926 55.48650 61.71203
21 42.79211 36.56486 49.01936
22 62.21935 58.57081 65.86788
23 62.90263 59.84986 65.95541
24 45.93016 38.03274 53.82758
25 54.75804 51.18143 58.33465
26 72.72682 69.44969 76.00395
27 73.76290 69.02088 78.50492
28 56.05502 52.44080 59.66924
29 79.56450 74.95247 84.17652
30 75.09964 70.87840 79.32089
```

**Inference:**

1.The coefficient for HC is positive whereas for OL is negative , hence while calculating overall supervisor performance we can say that HC(Handled employee Complaints) is directly related and OL is inversely related.

2. The p-value is 3.174e-07 which is very low ( also when compared to previous model) hence we can say that the model is significant and changes in the value of variables are related to change in the Supervisor performance.

3. The residuals show there is a linear relationship although when we have less predictor variables, the model shows more linearity.

**Comparison of Full Model and Reduced Model1**

Lets use anova test to compare full model and reduced model1.

```
> anova(suplm1,suplm2)
Analysis of Variance Table

Model 1: sup$SP ~ sup$HC + sup$NSP + sup$OL + sup$RBP + sup$CPP
+ sup$RA
Model 2: sup$SP ~ sup$HC + sup$OL
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     23 1149.0
2     27 1254.7 -4   -105.65 0.5287 0.7158
```

As we can see the p-value is significantly large we can say that no significant difference exists between the reduced model1 and Full model.

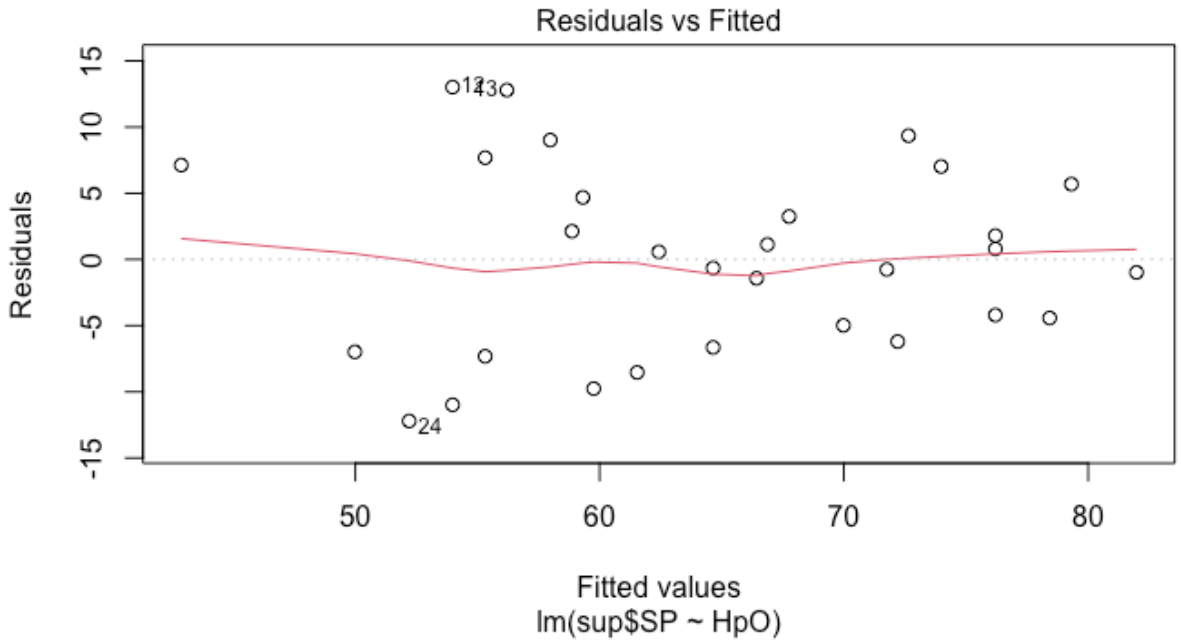
## 7.Reduced Model Two

Let us create the reduced model by using following variable:

```
> HpO= sup$HC+sup$OL
> suplm3 <- lm(sup$SP ~ HpO)
```

Let us create residual plot and we see that the model shows linearity.

```
> plot(suplm3,1))
```




Let's find the summary of the reduced model.

```
> summary(suplm3)

Call:
lm(formula = sup$SP ~ HpO)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2052  -5.8973  -0.0372   5.4364  13.0172

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.98821    7.38841   1.352    0.187
HpO          0.44439    0.05914   7.514 3.49e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 7.133 on 28 degrees of freedom
Multiple R-squared:  0.6685,    Adjusted R-squared:  0.6566
F-statistic: 56.46 on 1 and 28 DF,  p-value: 3.487e-08
```

The p-value is significantly small and lowest when compared to other models, hence the variables have significant differences in the means.

**Confidence Interval**

```
> confint(suplm3)
                 2.5 %     97.5 %
(Intercept) -5.1462575 25.1226846
HpO          0.3232388  0.5655406
```

**Prediction Interval**

The prediction interval is shown as below :

```
> summary(HpO)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   74.0   105.0   123.0   123.0   140.8   162.0

> newdata2 = data.frame(HpO=123)
> predict(suplm3, newdata2, interval="confidence")
       fit      lwr      upr
1 64.64815 61.98053 67.31576
```

Prediction interval lies between 61.98 and 67.315

**Comparison of Reduced Model 1 and Reduced Model2**

```
> > anova(suplm2,suplm3)
Analysis of Variance Table

Model 1: sup$SP ~ sup$HC + sup$OL
Model 2: sup$SP ~ HpO
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     27 1254.7
2     28 1424.6 -1   -169.95 3.6572 0.06649
```

When comparing the two reduced model we can say that value of p is big to conclude that no significant difference exist between the models.

## 8.Conclusion

We created various models such as with all variables and also reduced variables. When we created residual plots for all the models, they showed linearity. When we made comparison between the different models, it showed that no significant difference exists between them hence we would choose reduced model 2 with a smaller number of variables for the prediction of supervisor performance.