Biostatistics Problem ?

Done By:
Marc S Benison & Samia Porto

require{car}

From a survey of the clerical employees of a large financial organization, the data are aggregated from the questionnaires of the approximately 35 employees for each of 30 (randomly selected) departments. The numbers give the percent proportion of favorable responses to seven questions in each department.

A dataframe with 30 observations on 7 variables. The first column are the short names from the reference, the second one are the variable names in the data frame:

Rating is dependent –Y
The rest are independent—X

Rating →→ numeric Overall rating
Complaints →→ numeric Handling of employee complaints
Privileges →→ numeric Does not allow special privileges
Learning: →→ numeric Opportunity to learn
Raises →→ numeric Raises based on performances
Critical →→ numeric Too critical
Advancel →→ numeric Advancement

Question:

WHICH VARIABLES HAVE THE MOST EFFECT ON RATING??

attitude

```
> attitude
```

	rating	complaints	privileges	learning	raises	critical	advance
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
3	71	70	68	69	76	86	48
4	61	63	45	47	54	84	35
5	81	78	56	66	71	83	47
6	43	55	49	44	54	49	34
7	58	67	42	56	66	68	35
8	71	75	50	55	70	66	41
9	72	82	72	67	71	83	31
10	67	61	45	47	62	80	41
11	64	53	53	58	58	67	34
12	67	60	47	39	59	74	41
13	69	62	57	42	55	63	25
14	68	83	83	45	59	77	35
15	77	77	54	72	79	77	46
16	81	90	50	72	60	54	36
17	74	85	64	69	79	79	63
18	65	60	65	75	55	80	60
19	65	70	46	57	75	85	46
20	50	58	68	54	64	78	52
21	50	40	33	34	43	64	33
22	64	61	52	62	66	80	41
23	53	66	52	50	63	80	37
24	40	37	42	58	50	57	49
25	63	54	42	48	66	75	33
26	66	77	66	63	88	76	72
27	78	75	58	74	80	78	49
28	48	57	44	45	51	83	38
29	85	85	71	71	77	74	55
30	82	82	39	59	64	78	39

Before we begin working on the data, we need to attach the table above by using the function attach()

Then we'll set the variables as in X1,X2,X3..X_n

attach(attitude)

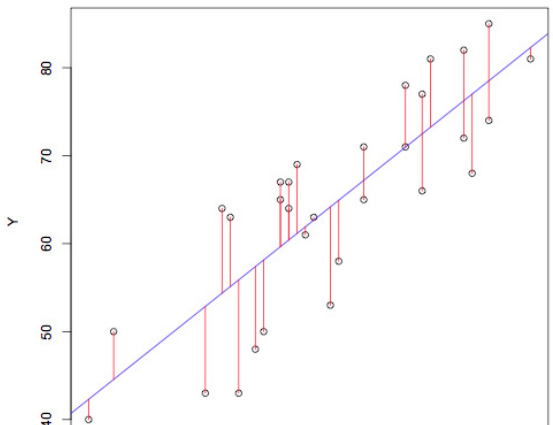
Y=rating
X1=complaints
X2=privileges
X3=learning
X4=raises
X5=critical
X6=advance

now we need to create a Full model

```
> FM=lm(Y~X1+X2+X3+X4+X5+X6)
> FM

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)

Coefficients:
(Intercept)           X1           X2           X3           X4
   10.78708      0.61319     -0.07305      0.32033      0.08173
         X5           X6
    0.03838     -0.21706
```

Simple linear regression

```
RM1=lm(Y~X1)
Yhat1=fitted(RM1)
plot(X1,Y)
```



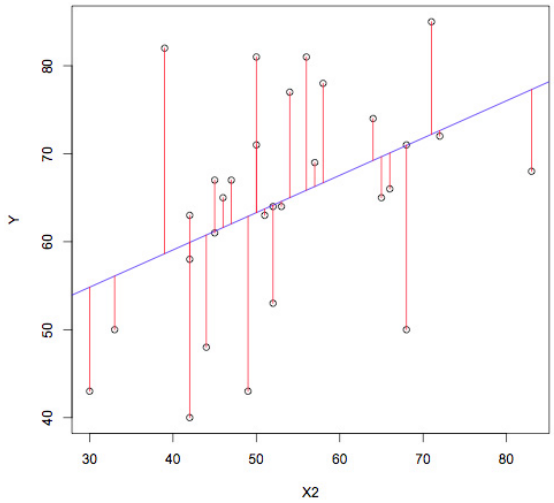
```
abline(RM1,col="blue")
segments(X1,Yhat1,X1,Y,col="red")
```

```
> anova(RM1)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
X1          1  2927.6 2927.58  59.861 1.988e-08 ***
Residuals  28  1369.4   48.91
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:

The graph and the p-value in the ANOVA function indicate that there is a relationship between the independent variable (X1) and the dependent variable (Y).

```
RM2=lm(Y~X2)
Yhat2=fitted(RM2)
plot(X2,Y)
abline(RM2,col="blue")
segments(X2,Yhat2,X2,Y,col="red")
```



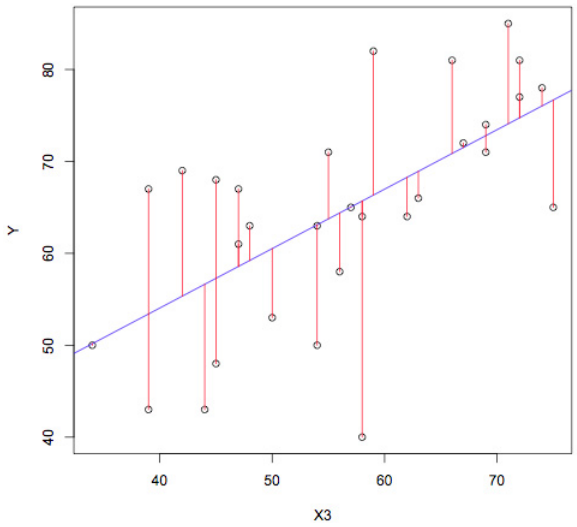

```
> anova(RM2)
```

```
Analysis of Variance Table
Response: Y
           Df  Sum Sq Mean Sq F value   Pr(>F)
X2          1   780.2  780.22  6.2121 0.01888 *
Residuals  28  3516.7  125.60
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:

The graph and the p-value in the ANOVA function indicate that there is no relationship between the independent variable (X2) and the dependent variable (Y).

```
RM3=lm(Y~X3)
Yhat3=fitted(RM3)
plot(X3,Y)
abline(RM3,col="blue")
segments(X3,Yhat3,X3,Y,col="red")
```

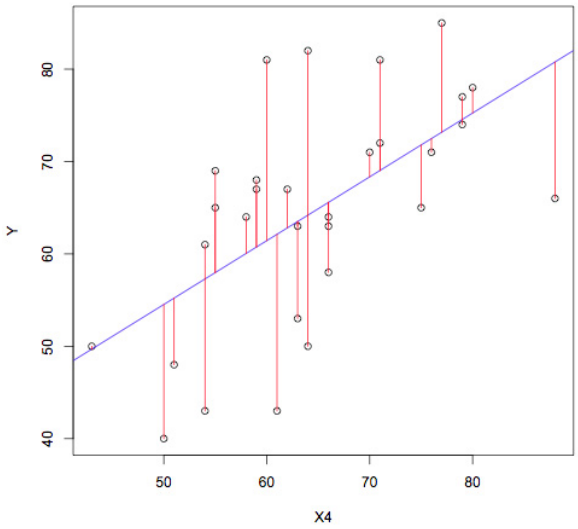



```
> anova(RM3)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
X3          1  1671.4 1671.41  17.825 0.0002311 ***
Residuals  28  2625.6   93.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:

The graph and the p-value in the ANOVA function indicate that there is a relationship between the independent variable (X3) and the dependent variable

```
RM4=lm(Y~X4)
Yhat4=fitted(RM4)
plot(X4,Y)
abline(RM4,col="blue")
segments(X4,Yhat4,X4,Y,col="red")
```




```
> anova(RM4)
Analysis of Variance Table
```

Response: Y

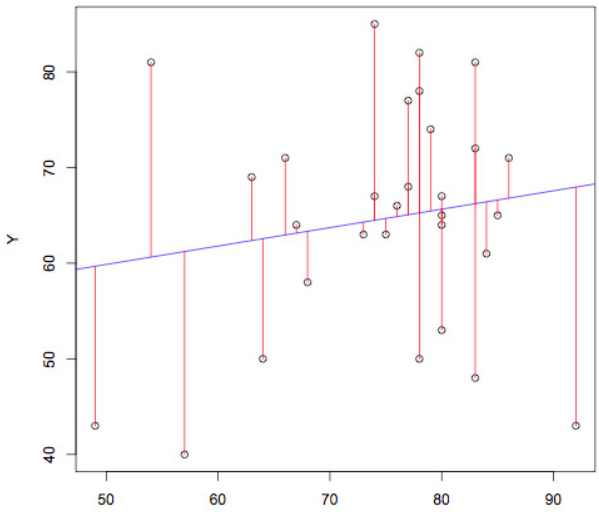
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X4	1	1496.5	1496.48	14.962	0.0005978	***
Residuals	28	2800.5	100.02			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Conclusion:

The graph and the p-value in the ANOVA function indicate that there is a relationship between the independent variable (X4) and the dependent variable

```
RM5=lm(Y~X5)
Yhat5=fitted(RM5)
plot(X5,Y)
abline(RM5,col="blue")
segments(X5,Yhat5,X5,Y,col="red")
```



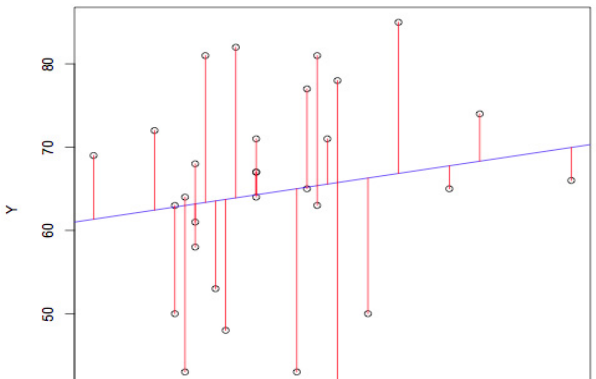

```
> anova(RM5)
Analysis of Variance Table

Response: Y
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X5	1	105.2	105.16	0.7024	0.4091
Residuals	28	4191.8	149.71		

Conclusion:

The graph and the p-value in the ANOVA function indicate that there is no relationship between the independent variable (X5) and the dependent variable

```
RM6=lm(Y~X6)
Yhat6=fitted(RM6)
plot(X6,Y)
abline(RM6,col="blue")
```



```
segments(X6,Yhat6,X6,Y,col="red")
```

```
> anova(RM6)
Analysis of Variance Table

Response: Y
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
X6          1   103.3   103.35     0.69  0.4132
Residuals 28  4193.6   149.77
```

Conclusion:

The graph and the p-value in the ANOVA function indicate that there is no relationship between the independent variable (X6) and the dependent variable

```
#Multipule regression full model
#Reduced model marginal

> FM=lm(Y~X1+X2+X3+X4+X5+X6)
> summary(FM)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)

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	
X1	0.61319	0.16098	3.809	0.000903	***
X2	-0.07305	0.13572	-0.538	0.595594	
X3	0.32033	0.16852	1.901	0.069925	.
X4	0.08173	0.22148	0.369	0.715480	
X5	0.03838	0.14700	0.261	0.796334	
X6	-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: 1.240e-05

Conclusion:
Intercept → p-value= 0.361634

X1 → p-value= 0.000903 → p-value < 0.01 → reject H_0, which shows that there is a relationship between the independent variable (X1) and the dependent variable (Y)

X2 → P-value= 0.595594 → p-value > 0.01 → accept H_0, which shows that there is no relationship between the independent variable (X2) and the dependent variable (Y)

X3 → P-value = 0.069925 → p-value > 0.01 → accept H_0, which shows that there is no relationship between the independent variable (X3) and the dependent variable (Y)

X4 → P-value = 0.715480→ p-value > 0.01 → accept H_0, which shows that there is no relationship between the independent variable (X4) and the dependent variable (Y)

X5 → P-value = 0.796334 → p-value > 0.01 → accept H_0, which shows that there is no relationship between the independent variable (X5) and the dependent variable (Y)

X6 → P-value = 0.235577 → p-value > 0.01 → accept H_0, which shows that there is no relationship between the independent variable (X6) and the dependent variable (Y)

Overall p-value = 1.240e-05 → we reject H_0 → this indicates that there is at least one independent variable that has a relationship with the dependent variable (Y).

```
> step(FM,direction="backward")

Start:  AIC=123.36
Y ~ X1 + X2 + X3 + X4 + X5 + X6

       Df  Sum of Sq    RSS    AIC
- X5    1      3.41  1152.4 121.45
- X4    1      6.80  1155.8 121.54
- X2    1     14.47  1163.5 121.74
- X6    1     74.11  1223.1 123.24
<none>              1149.0 123.36
- X3    1    180.50  1329.5 125.74
- X1    1    724.80  1873.8 136.04

Step:  AIC=121.45
Y ~ X1 + X2 + X3 + X4 + X6

       Df  Sum of Sq    RSS    AIC
- X4    1     10.61  1163.0 119.73
- X2    1     14.16  1166.6 119.82
- X6    1     71.27  1223.7 121.25
<none>              1152.4 121.45
- X3    1    177.74  1330.1 123.75
- X1    1    724.70  1877.1 134.09


Step:  AIC=119.73
Y ~ X1 + X2 + X3 + X6

       Df  Sum of Sq    RSS    AIC
- X2    1     16.10  1179.1 118.14
- X6    1     61.60  1224.6 119.28
<none>              1163.0 119.73
- X3    1    197.03  1360.0 122.42
- X1    1   1165.94  2328.9 138.56

Step:  AIC=118.14
Y ~ X1 + X3 + X6

       Df  Sum of Sq    RSS    AIC
- X6    1     75.54  1254.7 118.00
<none>              1179.1 118.14
- X3    1    186.12  1365.2 120.54
```

This function gives us the values of AIC for each of the independent variables as well as the linear model, which they make up.

We need to find the lowest AIC for the model.

The step function reduces the full model into multiple reduced model, by removing the independent variables with AIC that is lower than the model AIC. Each step removes the independent variable with the lowest AIC.. step by step until the AIC of the model is the lowest AIC value.

```
- X1     1     1259.91      2439.0   137.94

Step:  AIC=118
Y ~ X1 + X3

        Df   Sum of Sq    RSS      AIC
<none>                    1254.7   118.00
- X3     1     114.73     1369.4   118.63
- X1     1     1370.91    2625.6   138.16

Call:
lm(formula = Y ~ X1 + X3)

Coefficients:
(Intercept)         X1         X3
     9.8709       0.6435     0.2112
```

Conclusion:

The results tell us that the reduced model containing the independent variables X1 and X3 may be a better model to use than our full model.

```
# Comparison between the Full Model & the new reduced Model

> RM13=lm(Y~X1+X3)
> anova(RM13,FM)
Analysis of Variance Table

Model 1: Y ~ X1 + X3
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6
     Res.Df   RSS       Df     Sum of Sq        F        Pr(>F)
1     27        1254.7
2     23         1149.0     4         105.65      0.5287     0.7158
```

Conclusion:

The p-value= 0.7158 → P-value > 0.01 →we accept the H_0, which means that the reduced model is more preferred than the full model because it's the simplest model.

So back to our example:-

The results indicates that the variables that mostly affect the ratting are:

1) Complaints: Handling of employee complaint

2) Learning: Opportunity to learn