Statistical Methods II Project

Semester II

Name : Samahriti Mukherjee

Roll No.: BS2003

Indian statistical Institute

Contents

1 Introduction **3**

2 Description of the Data **3**

3 Importance of Analyse of the Data **3**

4 Analysis of the Data **4**
4.1 Regression Analysis and Testing of Hypothesis 4
4.2 Histogram 7
4.3 Q-Q Plot 7

5 Conclusion **8**

6 Acknowledgement **8**

1 Introduction

In this project, I will analyse a real life data (Productivity Prediction of Garment Employees Data Set) using some techniques which I have learnt in Statistics in the first two Methods courses.

2 Description of the Data

The file data.txt contains the data set which we are going to use for the project. I have collected the data set from here. This dataset includes important attributes of the garment manufacturing process and the productivity of the employees which had been collected manually and also been validated by the industry experts. The data set has 506 members with 14 attributes. The attributes which are used in the project are listed below :

S.NO.	Description	Variable	Type
1	Targeted productivity set by the Authority for each team for each day	targeted productivity	Continuous
2	Standard Minute Value, allocated time for a task	smv	Continuous
3	Work in progress, includes the number of unfinished items for products	wip	Discrete
4	amount of financial incentive (in BDT) that enables or motivates a particular course of action	incentive	Discrete
5	The amount of time when the production was interrupted due to several reasons	idle time	Continuous
6	The number of workers who were idle due to production interruption	idle men	Discrete
7	The actual % of productivity that was delivered by the workers (ranges from 0-1)	actual productivity	Continuous
8	Number of changes in the style of a particular product	no. of style change	Discrete

3 Importance of Analyse of the Data

The Garment Industry is one of the key examples of the industrial globalization of this modern era. It is a highly labour-intensive industry with lots of manual processes. Satisfying the huge global demand for garment products is mostly dependent on the production and delivery performance of the employees in the garment manufacturing companies. So, it is highly desirable among the decision makers in the garments industry to track, analyse and predict the productivity performance of the working teams in their factories.

4 Analysis of the Data

4.1 Regression Analysis and Testing of Hypothesis

Let us consider the actual productivity as response variable, say y, and targeted productivity (x_1), allocated time for a task (x_2), work in progress (x_3), incentive (x_4) as regressor variables. Therefore, the regression model is :

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

Now using R, we get the regression equation:

$$Y = 0.11751 + 0.6136 x_1 - 0.002018 x_2 + 0.0000002103 x_3 + 0.003358 x_4$$

Required R Codes and Outputs:

```
data=read.csv(file.choose(),header=TRUE)
tar=data$targeted_productivity
smv=data$smv
wip=data$wip
inc=data$incentive
time=data$idle_time
men=data$idle_men
style=data$no_of_style_change
workers=data$no_of_workers
y=data$actual_productivity
result=lm(y~tar+smv+wip+inc)
summary(result)
Call:
lm(formula = y ~ tar + smv + wip + inc)
Residuals:
     Min       1Q   Median       3Q      Max
-0.34265 -0.01996  0.00170  0.03244  0.35344
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.751e-01  2.283e-02   7.669 5.94e-14 ***
tar          6.136e-01  3.124e-02  19.638  < 2e-16 ***
smv         -2.018e-03  4.008e-04  -5.036 6.10e-07 ***
wip          2.103e-07  1.537e-06   0.137    0.891
inc          3.358e-03  1.173e-04  28.625  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.07311 on 686 degrees of freedom
  (506 observations deleted due to missingness)
Multiple R-squared:  0.7782,    Adjusted R-squared:  0.7769
F-statistic: 601.7 on 4 and 686 DF,  p-value: < 2.2e-16
```

Now we will try to answer some questions:

- **Is there a linear relationship between actual productivity and targeted productivity, allocated time for a task, work in progress, incentive ?**
  We test the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against $H_1 : \beta_i \neq 0$ for atleast one of i=1,2,3,4. Now we can see that p-value is less than 2×10^{-16}, which is very less, which indicates that there is enough evidence to reject the null hypothesis at 0.01 level of significance. So There is a linear relationship between actual productivity and targeted productivity, allocated time for a task, work in progress, incentive.

- **How much is the Strength of the Linear Relationship ?**
  We know that R^2 is the fraction of the variability in the model that can be explained by the least square linear regression model. Now in our dataset, value of R^2 is 0.7782. So we can say that 78% of the variation in actual productivity is explained by the least square linear regression model.

Now some other variables in the dataset seem to have no linear relationship with y. We will test whether this statement is true or false.
So suppose idle time (x_5), idle men (x_6), no. of style change (x_7) are other regressor variables along with the previous one. Then the regression model will be :

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon$$

We see the anova table for both the previous model and the recent model.

Required R Codes and Outputs:

```
anova(result)
Analysis of Variance Table
Response: y
           Df Sum Sq Mean Sq  F value    Pr(>F)
tar         1 8.0533  8.0533 1506.703 < 2.2e-16 ***
smv         1 0.3175  0.3175   59.408 4.508e-14 ***
wip         1 0.1139  0.1139   21.314 4.654e-06 ***
inc         1 4.3797  4.3797  819.394 < 2.2e-16 ***
Residuals 686 3.6667  0.0053
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
result.1=lm(y~tar+smv+wip+inc+time+men+style)
anova(result.1)
Analysis of Variance Table
Response: y
           Df Sum Sq Mean Sq   F value    Pr(>F)
tar         1 8.0533  8.0533 1628.4679 < 2.2e-16 ***
```

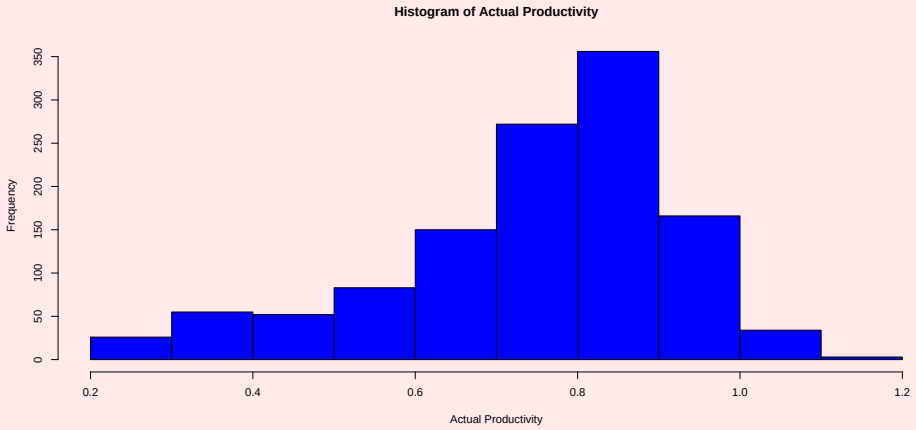
```
smv          1 0.3175  0.3175   64.2086 4.845e-15 ***
wip          1 0.1139  0.1139   23.0369 1.953e-06 ***

inc          1 4.3797  4.3797  885.6135 < 2.2e-16 ***
time         1 0.0058  0.0058    1.1748   0.27879
men          1 0.2590  0.2590   52.3728 1.238e-12 ***
style        1 0.0242  0.0242    4.8917   0.02732 *
Residuals  683 3.3777  0.0049
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
((3.6667-3.3777)/3)/(3.3777/683)
[1] 19.47943
qf(0.99,3,683)
[1] 3.810397
```

Now we pose a question that:

- **Is there a linear relationship between y and idle time (x_5), idle men (x_6), no. of style change (x_7)?**
  We test the hypothesis $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$ against $H_1 : \beta_i \neq 0$ for atleast one of i=5,6,7. We can see that value of our F statistic is 19.47943, with 3 and 683 degrees of freedom. Value of the F statistic at 0.01 level of significance is 3.810397, with 3 and 683 degrees of freedom, which is less than our observed F statistic value, so we can say that there is enough evidence to reject the null hypothesis at 0.01 level of significance, so there is a linear relationship between y and idle time (x_5), idle men (x_6), no. of style change (x_7).

4.2 Histogram

Now we look at the histogram of the actual productivity:



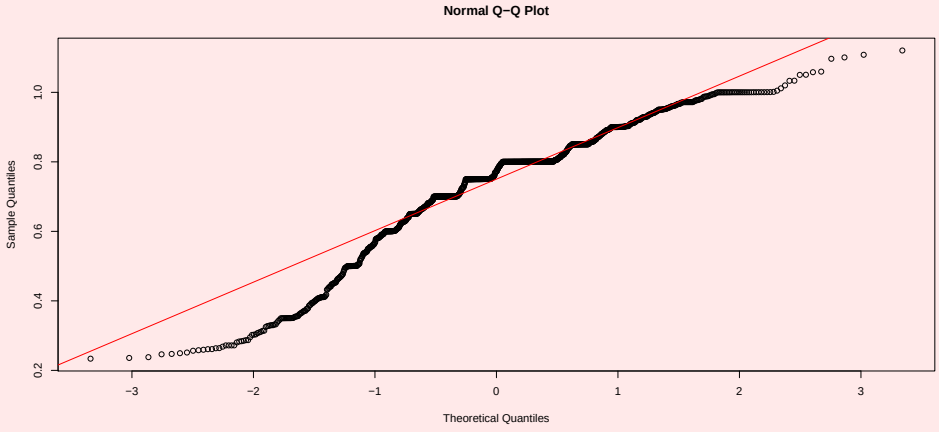
We can see that the histogram is negatively skewed, and when actual productivity is around 0.8, it has the highest frequency.

Required R Codes:

```
hist(y,xlab="Actual Productivity",col="Blue",main="Histogram of Actual Productivity")
```

4.3 Q-Q Plot

Now we try to see whether the dataset follows normal distribution or not. So we plot quantiles of the theoretical normal distribution on the X-axis and quantiles of the data on the Y-axis.



We can see that the Q-Q plot is not a straight-line. So we can conclude that the dataset does not follow normal distribution.

Required R Codes:

```
qqnorm(y)
qqline(y,col='red')
```

5 Conclusion

By analysing the dataset, we can conclude that this data set can be used to predict the actual productivity range (0-1) in advance. Before actual production, We will see whether the actual productivity range is low or high when we know the values of other regressor variables (targeted productivity, allocated time for a task, work in progress, incentive) and if the range is low then we will stop the production and will try to make some changes in the regressor variables. Values of idle time, idle men, no. of style change also affect the value of actual productivity . Also the value of actual productivity is concentrated much around 0.8. The distribution function of actual productivity does not follow a normal distribution.

6 Acknowledgement

I would like to express my special thanks of gratitude to my Statistics teacher Dr. Ayanendranath Basu, who gave me an opportunity to work on this project. It helped me a lot to learn about real life application of statistics and I came to know about a lot of new things. I am really thankful to my teacher.