BADM 575 Assignment 1

Xiying Zhao (xiyingz2)

2023-02-26

Load Packages and Data

```
d <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/00597/garments_worker_productiv
= TRUE, as.is = TRUE)
```

1. Data Processing

a. remove column wip

```
d$wip <- NULL
```

b. Create another variable named 'log_productivity'

```
d$log_productivity <- log(d$actual_productivity * 100)
```

c. create variable named "log_no_of_workers"

```
d$log_no_of_workers <- log(d$no_of_workers)
```

d.Convert the following variables to factor variables team, quarter, department, and day.

```
d$day <- as.factor(d$day)
d$quarter <- as.factor(d$quarter)
d$department <- as.factor(d$department)
d$team <- as.factor(d$team)
```

e. Create another variable called 'percentage_achivement'

```
d$percentage_achievment <- (d$actual_productivity - d$targeted_productivity) / d$targeted_productivity
```

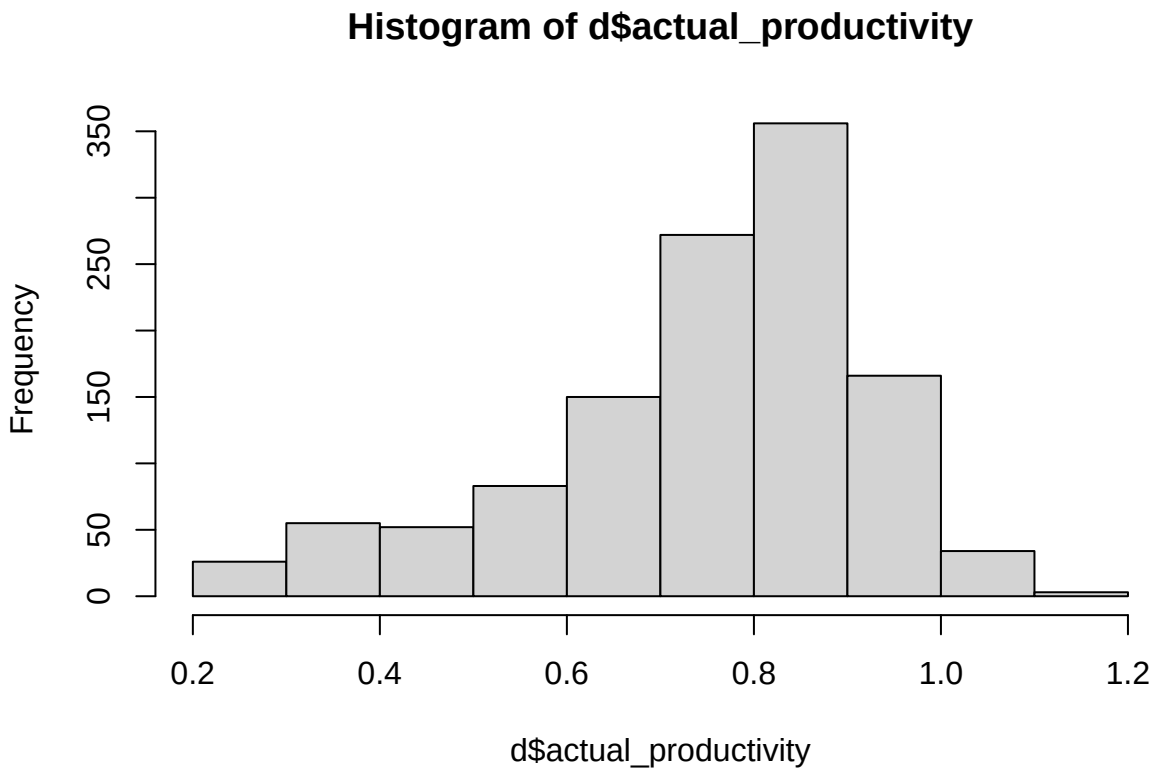
f. clean department variable

```
levels(d$department)<-c("finishing", "finishing", "sewing")
```

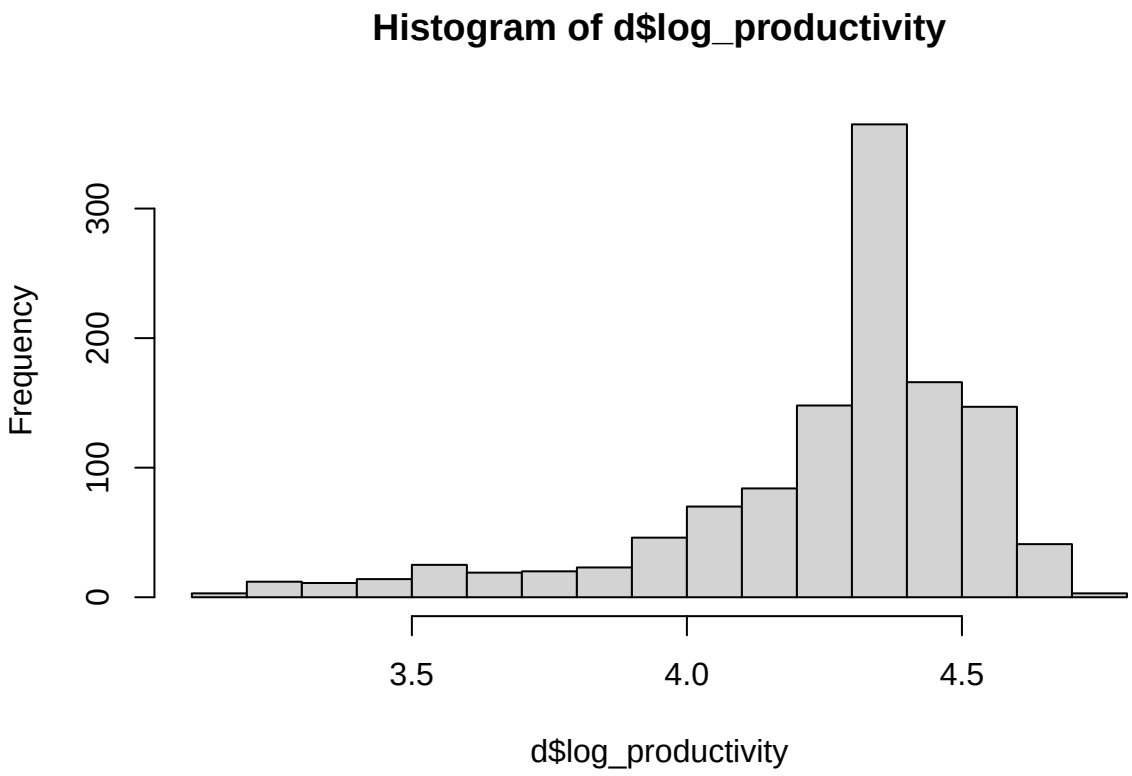
2.Exploratory Analysis

a. Create the histograms of actual_productivity and log_productivity. How does the distribution of log_productivity change with respect to actual_productivity? Do the same for number of workers.

```
# histogram of actual_productivity
hist(d$actual_productivity)
```

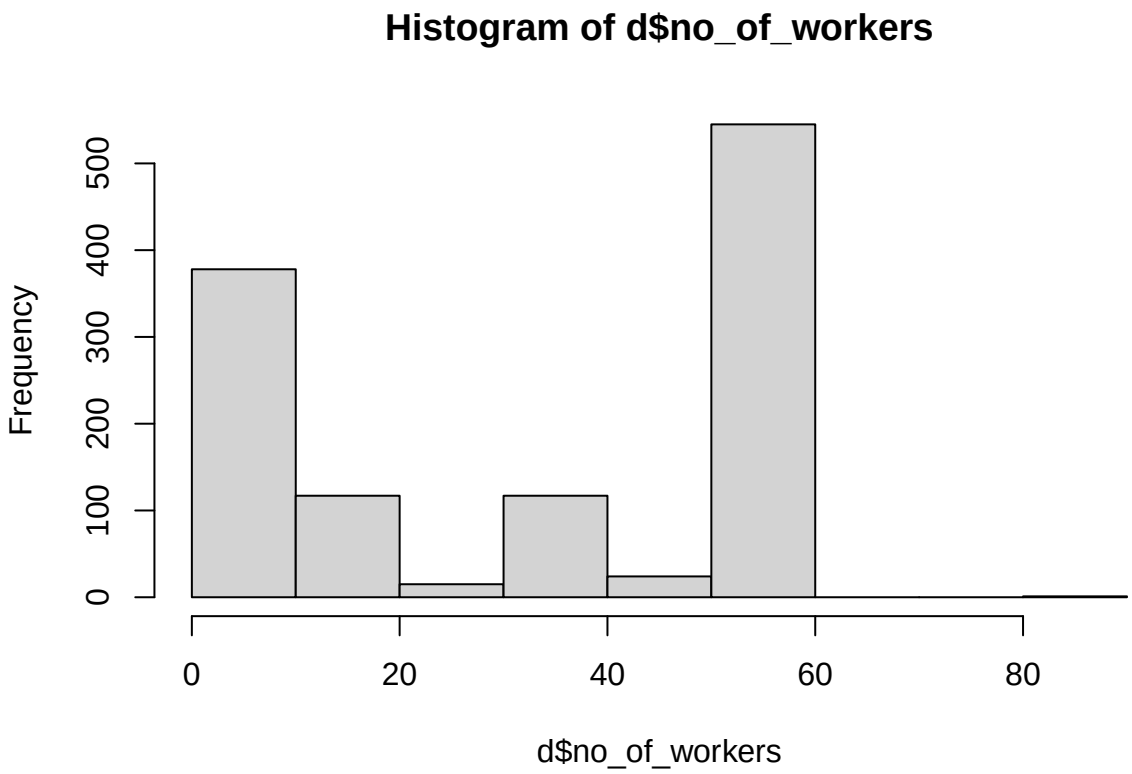



```
# histogram of log_productivity
hist(d$log_productivity)
```

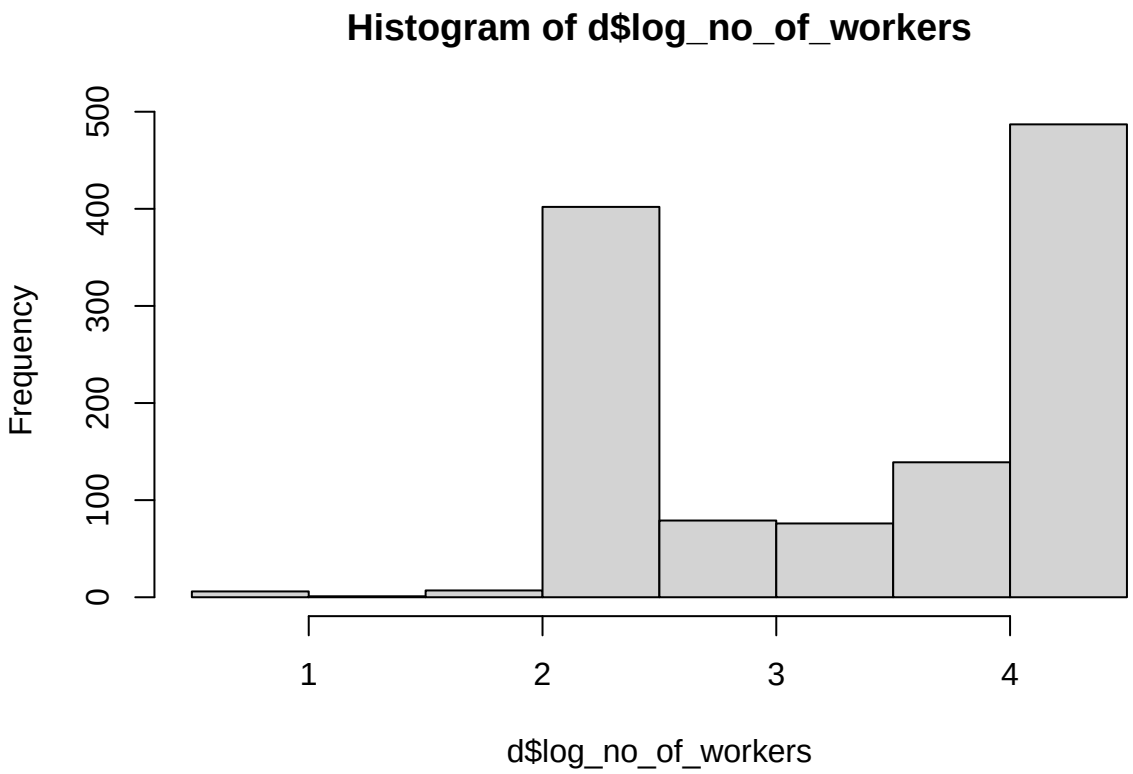


Observations: Comparing these two histograms, the distribution of log_productivity is more left skewed while the distribution of actual_productivity is more uniform.

```
# histogram of no_of_workers
hist(d$no_of_workers)
```



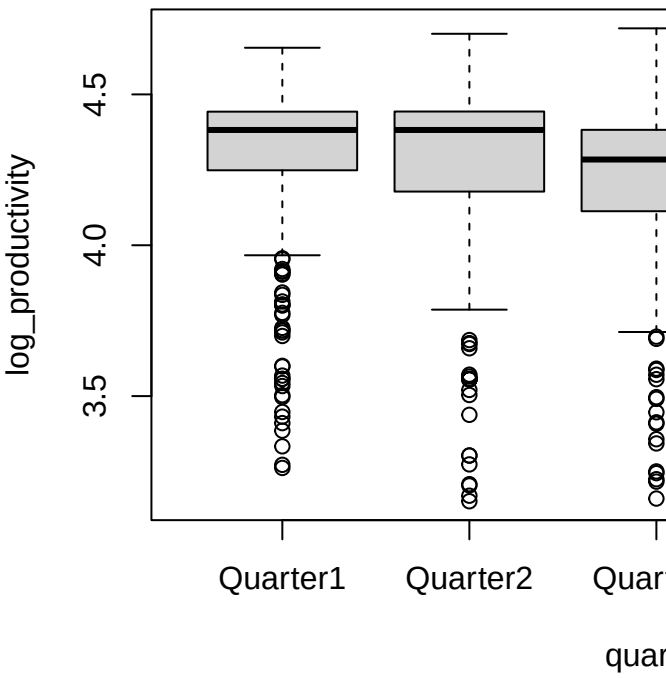

```
# histogram of log_no_of_workers
hist(d$log_no_of_workers)
```



Observations: Comparing these two histograms, the distribution of log_no_of_workers is more left-skewed while the distribution of no_of_workers is more right-skewed.

b. Each month is divided into five quarters, where approximately each week is a quarter. How does the distribution of logarithm of productivity change in each quarter?

```
boxplot(log_productivity~quarter, data=d)
```



Create a box plot of logarithm of productivity by quarter

Observations: The worker productivity increase towards the end of the month (quarter 5) as compared to other quarters. We can see from the box plot that the distribution of log_productivity is higher than the 4 quarters before.

Perform a t-test for quarter 5 with respect to (individually) all other quarters. Hypothesis statement: The mean of worker productivity stay the same at the end of the month (quarter 5) as compared to other quarters.

The mean of log_productivity of quarter 1-5 is:

```
#Average log_productivity of quarter 1-5
h1 = mean(d$log_productivity[d$quarter == 'Quarter1'])
h2 = mean(d$log_productivity[d$quarter == 'Quarter2'])
h3 = mean(d$log_productivity[d$quarter == 'Quarter3'])
h4 = mean(d$log_productivity[d$quarter == 'Quarter4'])
h5 = mean(d$log_productivity[d$quarter == 'Quarter5'])
print(c(h1,h2,h3,h4,h5))
```

```
## [1] 4.290405 4.274406 4.215843 4.218121 4.381652
```

The Std. dev. of log_productivity of quarter 1-5 is

```
#Std. dev. log_productivity of quarter 1-5
s1 = sd(d$log_productivity[d$quarter == 'Quarter1'])
s2 = sd(d$log_productivity[d$quarter == 'Quarter2'])
```

```
s3 = sd(d$log_productivity[d$quarter == 'Quarter3'])
s4 = sd(d$log_productivity[d$quarter == 'Quarter4'])
s5 = sd(d$log_productivity[d$quarter == 'Quarter5'])
print(c(s1,s2,s3,s4,s5))
```

```
## [1] 0.2591165 0.2852128 0.3028682 0.3145746 0.2800994
```

The number of log_productivity of quarter 1-5 is

```
# number of log_productivity of quarter 1-5
n1=sum(d$quarter == 'Quarter1')
n2=sum(d$quarter == 'Quarter2')
n3=sum(d$quarter == 'Quarter3')
n4=sum(d$quarter == 'Quarter4')
n5=sum(d$quarter == 'Quarter5')
print(c(n1,n2,n3,n4,n5))
```

```
## [1] 360 335 210 248  44
```

$$H_0 : h_i = h_5 H_a : h_i \neq h_5 (i = 1, 2, 3, 4)$$

The t-test statistic in this case

$$t = \frac{\hat{h}_i - \hat{h}_5}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_5^2}{n_5}}} (i = 1, 2, 3, 4)$$

```
# t-test-1
t1=t.test(d$log_productivity[d$quarter == 'Quarter1'], d$log_productivity[d$quarter == 'Quarter5'])
t1
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$quarter == "Quarter1"] and d$log_productivity[d$quarter == "Quarter5"]
## t = -2.056, df = 52.397, p-value = 0.04478
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.180285764 -0.002207791
## sample estimates:
## mean of x mean of y
##  4.290405  4.381652
```

In t-test-1, the mean of log_productivity of quarter 1 is 4.290405 compared to 4.381652 of quarter 5. And the p-value is 0.04478(less than 0.05) which means that we can reject the null hypothesis with 95% confidence.

```
# t-test-2
t2=t.test(d$log_productivity[d$quarter == 'Quarter2'], d$log_productivity[d$quarter == 'Quarter5'])
t2
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$quarter == "Quarter2"] and d$log_productivity[d$quarter == "Quarter5"]
## t = -2.3827, df = 55.377, p-value = 0.02064
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1974338 -0.0170569
## sample estimates:
## mean of x mean of y
##  4.274406  4.381652
```

In t-test-2, the mean of log_productivity of quarter 2 is 4.274406 compared to 4.381652 of quarter 5. And the p-value is 0.02064(less than 0.05) which means that we can reject the null hypothesis with 95% confidence.

```
# t-test-3
t3=t.test(d$log_productivity[d$quarter == 'Quarter3'], d$log_productivity[d$quarter == 'Quarter5'])
t3
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$quarter == "Quarter3"] and d$log_productivity[d$quarter == "Quarter5"]
## t = -3.5192, df = 65.835, p-value = 0.0007905
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2598828 -0.0717350
## sample estimates:
## mean of x mean of y
##  4.215843  4.381652
```

In t-test-3, the mean of log_productivity of quarter 3 is 4.215843 compared to 4.381652 of quarter 5. And the p-value is 0.0007905(less than 0.05) which means that we can reject the null hypothesis with 95% confidence.

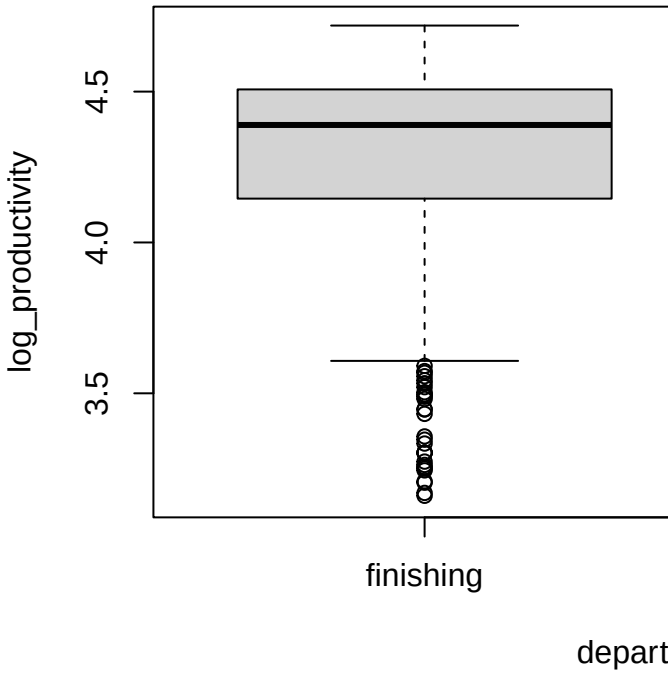
```
# t-test-4
t4=t.test(d$log_productivity[d$quarter == 'Quarter4'], d$log_productivity[d$quarter == 'Quarter5'])
t4
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$quarter == "Quarter4"] and d$log_productivity[d$quarter == "Quarter5"]
## t = -3.5008, df = 63.842, p-value = 0.0008519
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.25685499 -0.07020621
## sample estimates:
## mean of x mean of y
##  4.218121  4.381652
```

In t-test-4, the mean of log_productivity of quarter 4 is 4.218121 compared to 4.381652 of quarter 5. And the p-value is 0.0008519(less than 0.05) which means that we can reject the null hypothesis with 95% confidence.

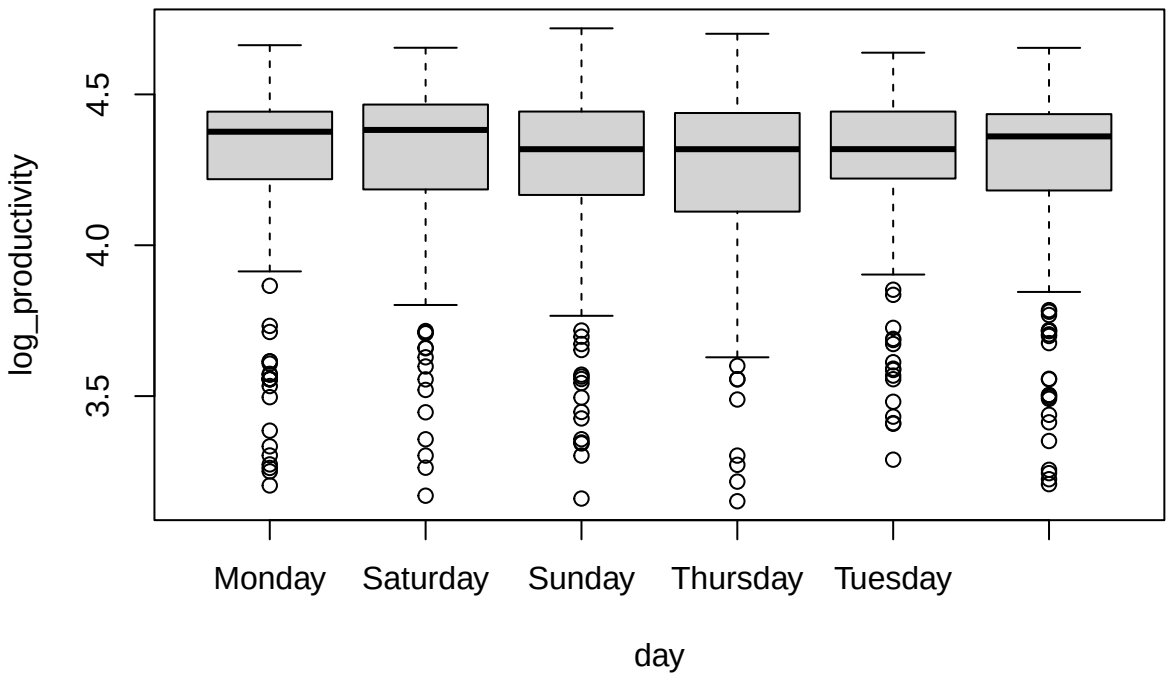
c. Repeat part (b) for department instead of quarter, day instead of quarter, and no_of_style_change instead of quarter. In these cases, perform the t-test for all pairs of departments and all pairs of style changes. For day, compare Sunday with all other weekdays.

```
# department instead of quarter
boxplot(log_productivity~department, data=d)
```

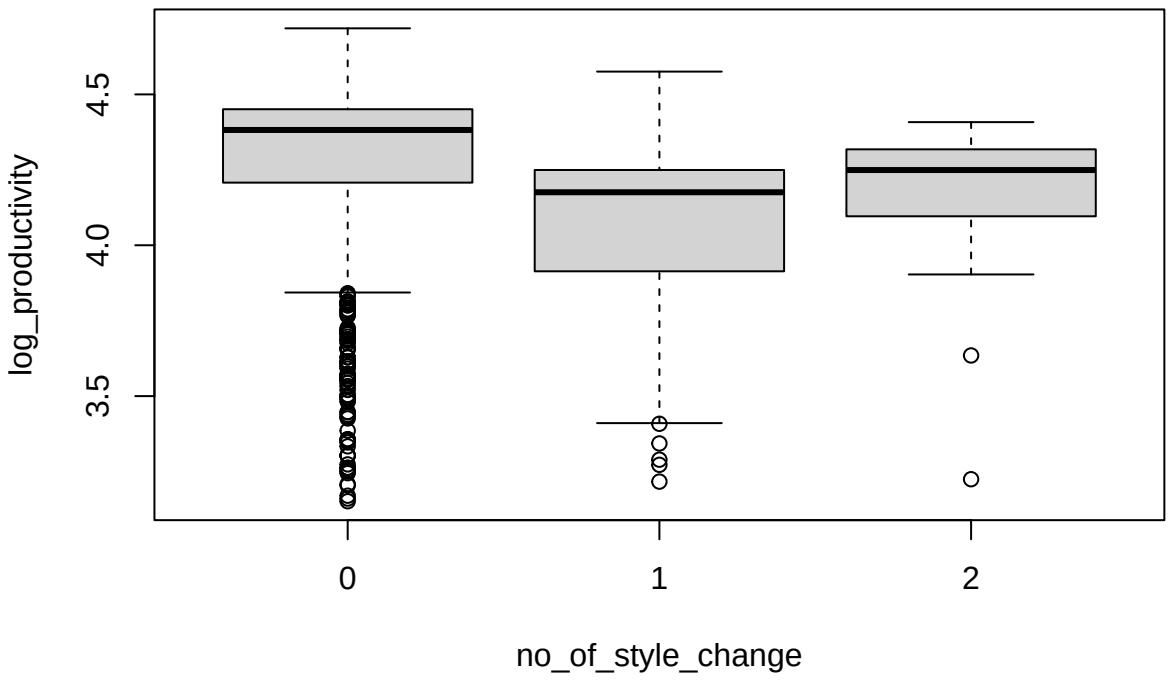


Create a box plot of logarithm of productivity by quarter

```
# day instead of quarter
boxplot(log_productivity~day, data=d)
```




```
# no_of_style_change instead of quarter
boxplot(log_productivity~no_of_style_change, data=d)
```



Perform a t-test. t-test for all pairs of departments

```
t.department=t.test(log_productivity~department,data=d)
t.department
```

```
##
##  Welch Two Sample t-test
##
## data:  log_productivity by department
## t = 1.5098, df = 941.23, p-value = 0.1314
## alternative hypothesis: true difference in means between group finishing and group sewing is not equa
## 95 percent confidence interval:
##  -0.007920583  0.060749866
## sample estimates:
## mean in group finishing    mean in group sewing
##                4.276473                4.250058
```

In t-test for all pairs of departments, the mean of log_productivity of group finishing is 4.276473 compared to 4.250058 of group sewing . And the p-value is 0.1314(greater than 0.05) which means that we cannot reject the null hypothesis with 95% confidence.

t-test for all pairs of style changes

```r
# no_of_style_change == 0 v.s. no_of_style_change == 1
t.style01=t.test(d$log_productivity[d$no_of_style_change == 0], d$log_productivity[d$no_of_style_change
t.style01
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$no_of_style_change == 0] and d$log_productivity[d$no_of_style_change == 1
## t = 6.9316, df = 135.33, p-value = 1.547e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1464295 0.2633414
## sample estimates:
## mean of x mean of y
##  4.283915  4.079030
```

```r
# no_of_style_change == 0 v.s. no_of_style_change == 2
t.style02=t.test(d$log_productivity[d$no_of_style_change == 0], d$log_productivity[d$no_of_style_change
t.style02
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$no_of_style_change == 0] and d$log_productivity[d$no_of_style_change == 2
## t = 2.6885, df = 34.807, p-value = 0.01094
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0282116 0.2023328
## sample estimates:
## mean of x mean of y
##  4.283915  4.168643
```

```r
# no_of_style_change == 1 v.s. no_of_style_change == 2
t.style12=t.test(d$log_productivity[d$no_of_style_change == 1], d$log_productivity[d$no_of_style_change
t.style12
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$no_of_style_change == 1] and d$log_productivity[d$no_of_style_change == 2
## t = -1.7709, df = 63.827, p-value = 0.08135
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.19070858  0.01148206
## sample estimates:
## mean of x mean of y
##  4.079030  4.168643
```

In t-test for all pairs of styles, the mean of log_productivity of three groups are 4.283915, 4.079030, 4.168643. And the p-value is 1.547e-10 for 0&1, 0.01094 for 0&2, 0.08135 for 1&2, among which the former two are less than 0.05, which means that we can reject the hypothesis that the mean of group0 is the same with group1 or group2 with 95% confidence.

t-test for day, compare Sunday with all other weekdays.

```
t.day1=t.test(d$log_productivity[d$day == "Monday"], d$log_productivity[d$day == "Sunday"])
t.day1
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$day == "Monday"] and d$log_productivity[d$day == "Sunday"]
## t = 0.25071, df = 398.5, p-value = 0.8022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05141280  0.06644255
## sample estimates:
## mean of x mean of y
##  4.258583  4.251068
```

```
t.day2=t.test(d$log_productivity[d$day == "Tuesday"], d$log_productivity[d$day == "Sunday"])
t.day2
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$day == "Tuesday"] and d$log_productivity[d$day == "Sunday"]
## t = 0.95673, df = 397.74, p-value = 0.3393
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02798409  0.08104195
## sample estimates:
## mean of x mean of y
##  4.277597  4.251068
```

```
t.day3=t.test(d$log_productivity[d$day == "Wednesday"], d$log_productivity[d$day == "Sunday"])
t.day3
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$day == "Wednesday"] and d$log_productivity[d$day == "Sunday"]
## t = 0.030093, df = 408.99, p-value = 0.976
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05697791  0.05874953
## sample estimates:
## mean of x mean of y
##  4.251953  4.251068
```

```
t.day4=t.test(d$log_productivity[d$day == "Thursday"], d$log_productivity[d$day == "Sunday"])
t.day4
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$day == "Thursday"] and d$log_productivity[d$day == "Sunday"]
## t = -0.20894, df = 399.88, p-value = 0.8346
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06264306  0.05060680
## sample estimates:
## mean of x mean of y
##  4.245050  4.251068
```

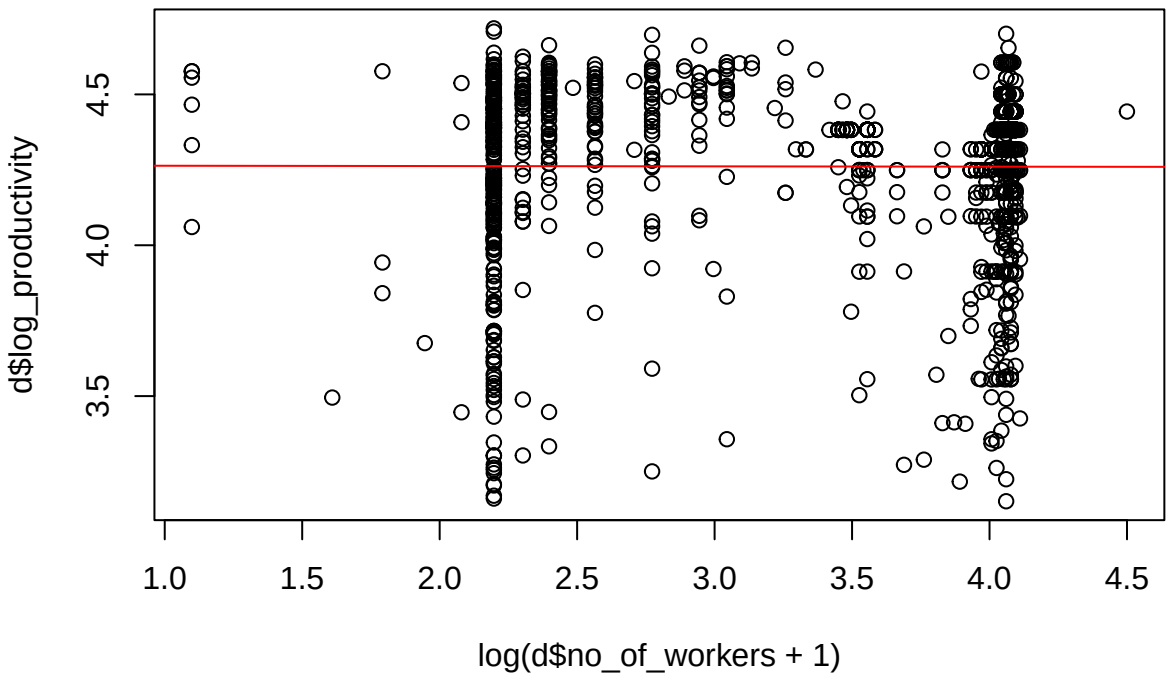
```
t.day6=t.test(d$log_productivity[d$day == "Saturday"], d$log_productivity[d$day == "Sunday"])
t.day6
```

```
##
##  Welch Two Sample t-test
##
## data:  d$log_productivity[d$day == "Saturday"] and d$log_productivity[d$day == "Sunday"]
## t = 1.1549, df = 386.9, p-value = 0.2489
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02382729  0.09166755
## sample estimates:
## mean of x mean of y
##  4.284988  4.251068
```

In t-test for all days, the p-value is all greater than 0.05, which means that we cannot reject the hypothesis that the mean of Sunday is the same with other weekdays with 95% confidence.

d. Perform a scatter plot of the natural logarithm of no_of_workers +1 on x–axis and natural logarithm of productivity on y-axis. What do you observe? Comment on any pattern that you may observe. Report the correlation coefficient between the two variables.

```
plot(log(d$no_of_workers+1), d$log_productivity)
abline(lm(d$log_productivity ~ log(d$no_of_workers+1)),col="red")
```



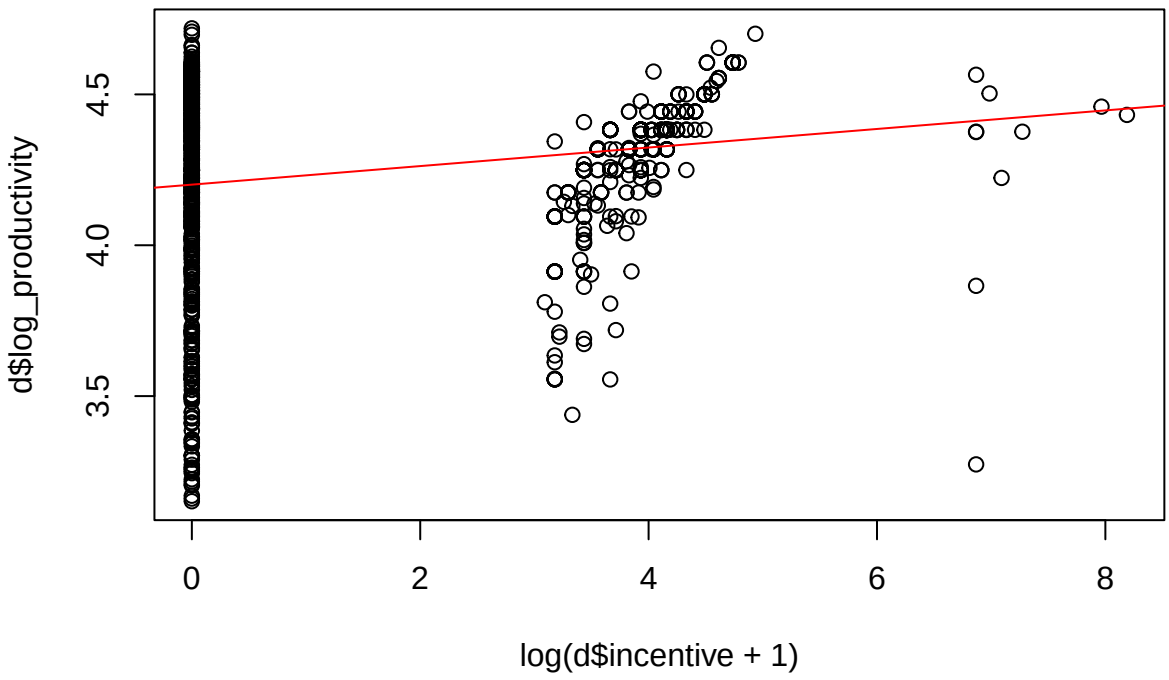

```
cor(log(d$no_of_workers+1), d$log_productivity)
```

```
## [1] -0.00286047
```

Observations: We can see that the the natural logarithm of no_of_workers +1 sightly influence the natural logarithm of productivity, and the coefficient is -0.00286047.

e. Perform a scatter plot of the natural logarithm of incentive + 1 on x-axis and natural logarithm of productivity on y-axis. What do you observe? Comment on any patterns that you may observe. Report the correlation coefficient between the two variables.

```
plot(log(d$incentive+1), d$log_productivity)
abline(lm(d$log_productivity~log(d$incentive+1)),col="red")
```



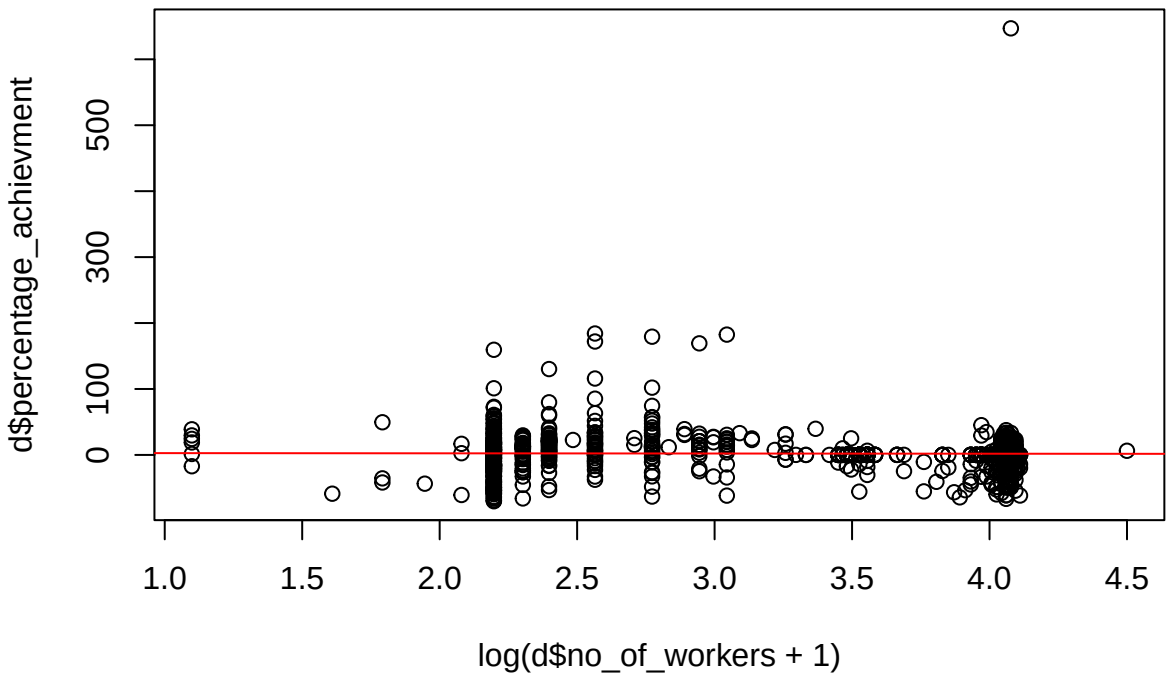

```
cor(log(d$incentive+1), d$log_productivity)
```

```
## [1] 0.2149898
```

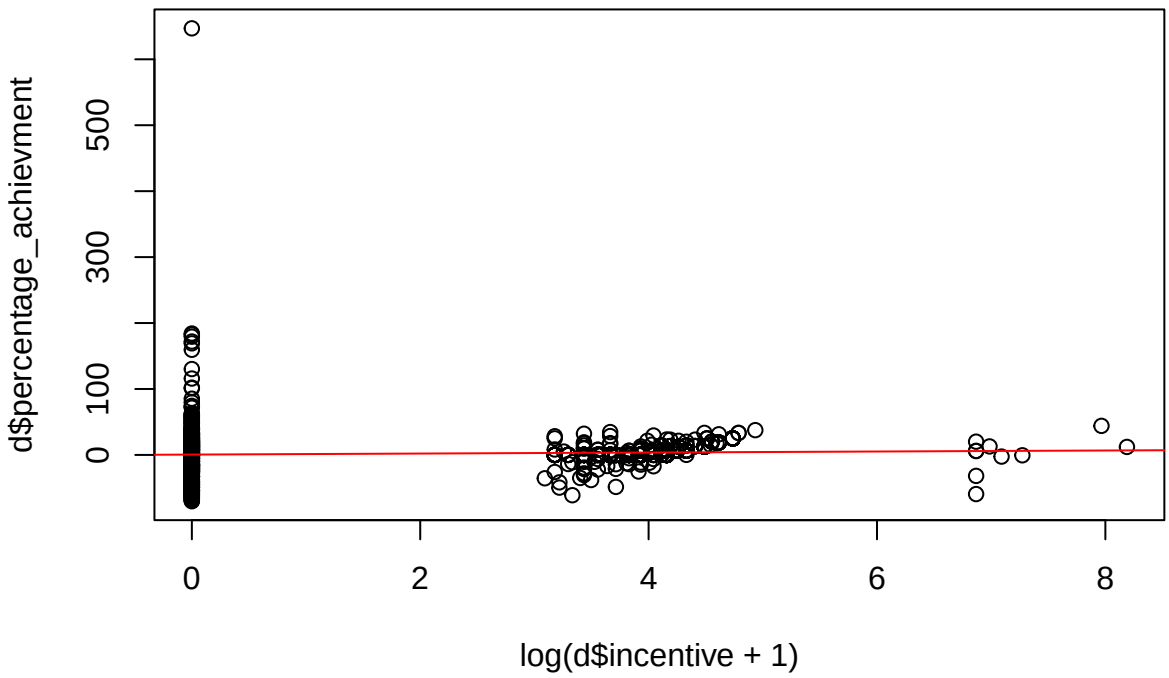
Observations: We can see a slightly positive linear correlation between the natural logarithm of incentive + 1 and natural logarithm of productivity, and the coefficient is 0.2149898.

f. Repeat (d) and (e) for percentage_achievement instead of logarithm of productivity.

```
plot(log(d$no_of_workers+1), d$percentage_achievment)
abline(lm(d$percentage_achievment~log(d$no_of_workers+1)),col="red")
```




```
plot(log(d$incentive+1),d$percentage_achievment)
abline(lm(d$percentage_achievment~log(d$incentive+1)),col="red")
```



2. Regression Analysis ### a. Estimate an ordinary least square repression (OLS) with natural logarithm of productivity as response variable and natural logarithm of no_of_workers + 1 as the predictor variable. Comment on the relationship between the response and the predictor variable. Is team size (number of workers in a team) a good predictor of productivity? Does this finding conform to the exploratory analysis in 2(d)? What is the estimated regression equation? How much of the variance in the response is explained by the predictor? (Comment on the R-square, the intercept, slope, and the t-statistics of the intercept and slope, and the p-values). Finally, plot the regression equation on the scatterplot of the predictor and response.

```
model1 <- lm(log_productivity~log(no_of_workers+1),data=d)
summary(model1)
```

```
##
## Call:
## lm(formula = log_productivity ~ log(no_of_workers + 1), data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10897 -0.08560  0.08581  0.18249  0.45658
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.2645096  0.0342612 124.471   <2e-16 ***
## log(no_of_workers + 1) -0.0009996  0.0101086  -0.099    0.921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

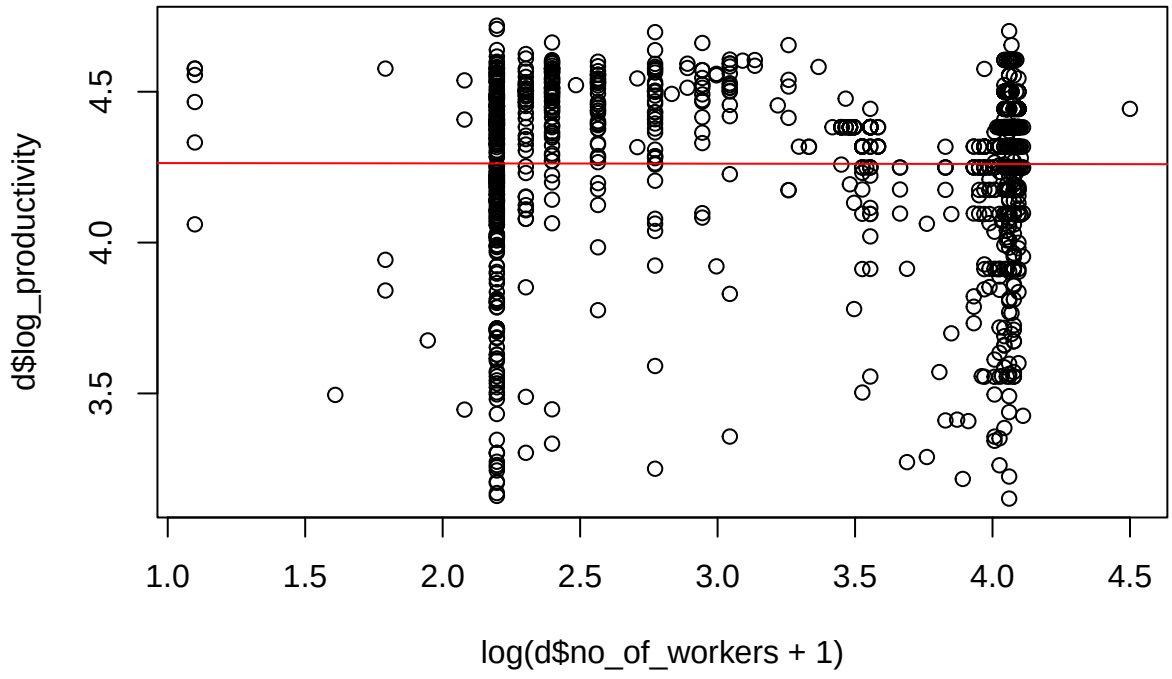
```
##
## Residual standard error: 0.2895 on 1195 degrees of freedom
## Multiple R-squared:  8.182e-06,  Adjusted R-squared:  -0.0008286
## F-statistic: 0.009778 on 1 and 1195 DF,  p-value: 0.9212
```

We can see that the correlation between these two variables is not significant, therefore, the team size (number of workers in a team) may not be a good predictor of productivity. This finding conforms to the exploratory analysis in 2(d). The estimated regression equation is

$$E[log(productivity)|log(no\_of\_workers + 1)] = 4.2645096 - 0.0009996 \times log(no\_of\_workers + 1).$$

Since the Multiple R-squared: 8.182e-06 is close to 0, we can say that nearly none of the variance in the response is explained by the predictor.

```
plot(log(d$no_of_workers+1), d$log_productivity)
abline(lm(log_productivity ~ log(no_of_workers + 1), data = d), col="red")
```



b. Repeat (a) with logarithm of incentives + 1 as the predictor.

```
model2 <- lm(log_productivity~log(incentive+1),data=d)
summary(model2)
```

```
##
## Call:
## lm(formula = log_productivity ~ log(incentive + 1), data = d)
##
```

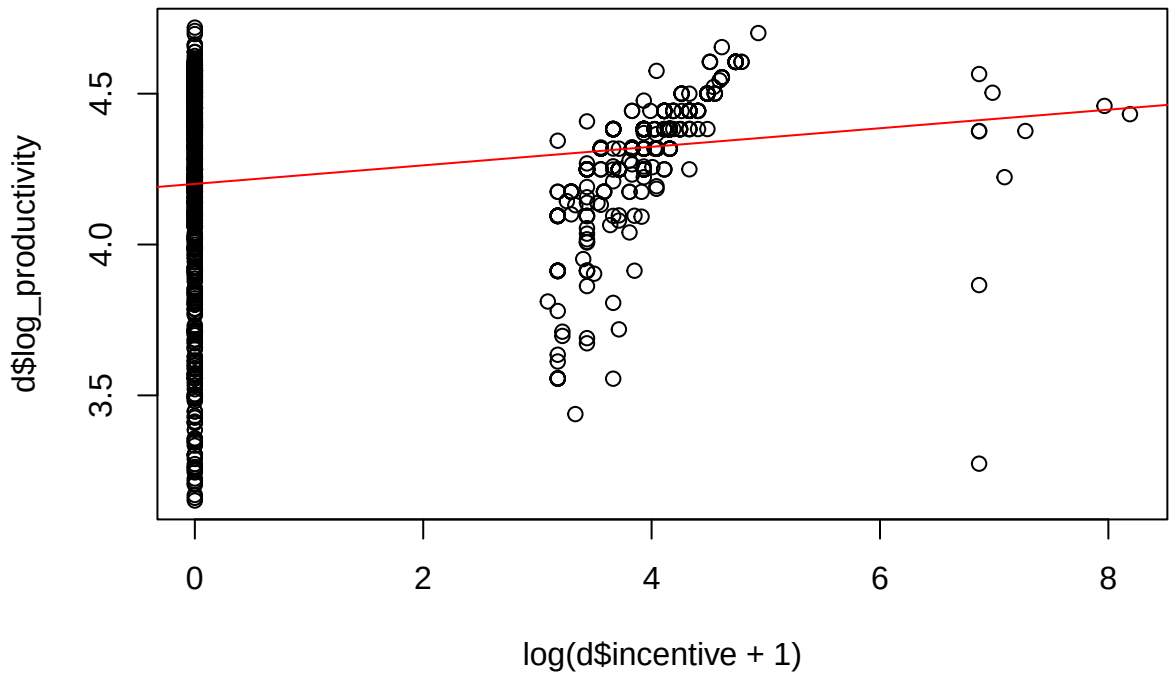
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13844 -0.07271  0.05377  0.17805  0.51804
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.200851   0.011390  368.83  < 2e-16 ***
## log(incentive + 1) 0.030749   0.004041    7.61 5.54e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2827 on 1195 degrees of freedom
## Multiple R-squared:  0.04622,    Adjusted R-squared:  0.04542
## F-statistic: 57.91 on 1 and 1195 DF,  p-value: 5.542e-14
```

We can see that the correlation between these two variables is significant according to the p-value<0.001, therefore, incentives can be a good predictor of productivity. The estimated regression equation is

$$E[log(productivity)|log(incentives + 1)] = 4.200851 + 0.030749 \times log(incentives + 1).$$

Since the Multiple R-squared is 0.04622, we can say that 4.62% of the variance in the response is explained by the predictor.

```
plot(log(d$incentive+1), d$log_productivity)
abline(lm(log_productivity ~ log(incentive + 1), data = d), col="red")
```



c. Estimate the regression equation for log of actual productivity as response and the following variables as predictors: log of no_of_workers + 1, log of incentive + 1, log of targeted productivity, no_of_style_change, quarter (factor variable), department (factor variable), day (factor variable) and team (factor variable). Show the regression summary. Answer the following question:

```
model3 <- lm(log_productivity~log(no_of_workers+1)+log(incentive + 1) + log(targeted_productivity) + no
summary(model3)
```

```
##
## Call:
## lm(formula = log_productivity ~ log(no_of_workers + 1) + log(incentive +
##     1) + log(targeted_productivity) + no_of_style_change + quarter +
##     department + day + team, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28802 -0.06598  0.03551  0.12945  1.08247
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.066351   0.083296  48.818  < 2e-16 ***
## log(no_of_workers + 1)       0.178241   0.031564   5.647 2.05e-08 ***
## log(incentive + 1)           0.064733   0.005933  10.910  < 2e-16 ***
## log(targeted_productivity)   0.493178   0.043217  11.412  < 2e-16 ***
## no_of_style_change          -0.029110   0.018925  -1.538 0.124288
## quarterQuarter2             -0.006396   0.018710  -0.342 0.732543
## quarterQuarter3             -0.036148   0.021455  -1.685 0.092289 .
## quarterQuarter4             -0.050495   0.020909  -2.415 0.015888 *
## quarterQuarter5              0.106628   0.040391   2.640 0.008404 **
## departmentsewing            -0.493515   0.056240  -8.775  < 2e-16 ***
## daySaturday                  0.030091   0.025406   1.184 0.236482
## daySunday                    0.014940   0.024420   0.612 0.540792
## dayThursday                  0.013366   0.024845   0.538 0.590711
## dayTuesday                   0.038951   0.024339   1.600 0.109791
## dayWednesday                 0.018706   0.024165   0.774 0.439030
## team2                       -0.050642   0.033376  -1.517 0.129462
## team3                       -0.014137   0.034684  -0.408 0.683641
## team4                       -0.037371   0.033936  -1.101 0.271035
## team5                       -0.059631   0.035197  -1.694 0.090491 .
## team6                       -0.089064   0.036177  -2.462 0.013964 *
## team7                       -0.127264   0.034826  -3.654 0.000269 ***
## team8                       -0.122107   0.033665  -3.627 0.000299 ***
## team9                       -0.106895   0.033693  -3.173 0.001550 **
## team10                      -0.121438   0.034062  -3.565 0.000378 ***
## team11                      -0.133209   0.035441  -3.759 0.000179 ***
## team12                      -0.007741   0.035454  -0.218 0.827193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2428 on 1171 degrees of freedom
## Multiple R-squared:  0.3106, Adjusted R-squared:  0.2959
## F-statistic: 21.11 on 25 and 1171 DF,  p-value: < 2.2e-16
```

i. Which of the following variables significantly affect worker productivity and which direction? State the level of significance. log(no_of_workers + 1), log(incentive + 1) ,log(targeted_productivity), department, team7-11 are significant at 99.9% level, among which the increase in log(no_of_workers + 1), log(incentive + 1) ,log(targeted_productivity) can cause an increase in worker productivity. And the sewing department has a lower worker productivity.

ii. On the average how much does log of productivity change with one incremental style change. Since the coefficient of no_of_style_change is -0.029110, productivity would decrease by 2.9% with one incremental style change since there is a log unit.

iii. What is the change in log of productivity for quarter 2, 3, 4 and 5 with respect to quarter 1. Which of these changes are statistically significant? Quarter 2,3, 4 have negative impact whereas quarter 5 has a positive impact on log of productivity with respect to quarter 1. Among these, quarter 4 and 5 are significant at 95% and 99% level respectively.

iv. How does the productivity of sewing department compare with the finishing department? The sewing department has a lower worker productivity because the coefficient of departmentsewing is -0.493515.

v. Write down the regression equation for the following cases:

1. Sewing department for a Sunday of quarter 4 for team 10.

```
4.066351 -0.050495 -0.493515+0.014940-0.121438
```

```
## [1] 3.415843
```

$E[log\_productivity|(department = Sewing, day = Sunday, quarter = 4, team = 10)] = 3.415843+0.178241 \times log(no\_of\_wo$

2. Finishing department for a Wednesday of quarter 1 for team 4.

```
4.066351+0+0.018706+0-0.037371
```

```
## [1] 4.047686
```

$E[log\_productivity|(department = Finishing, day = Wednesday, quarter = 1, team = 4)] = 4.047686+0.178241 \times log(no\_o$

3. Finishing department for a Monday of quarter 2 for team 8.

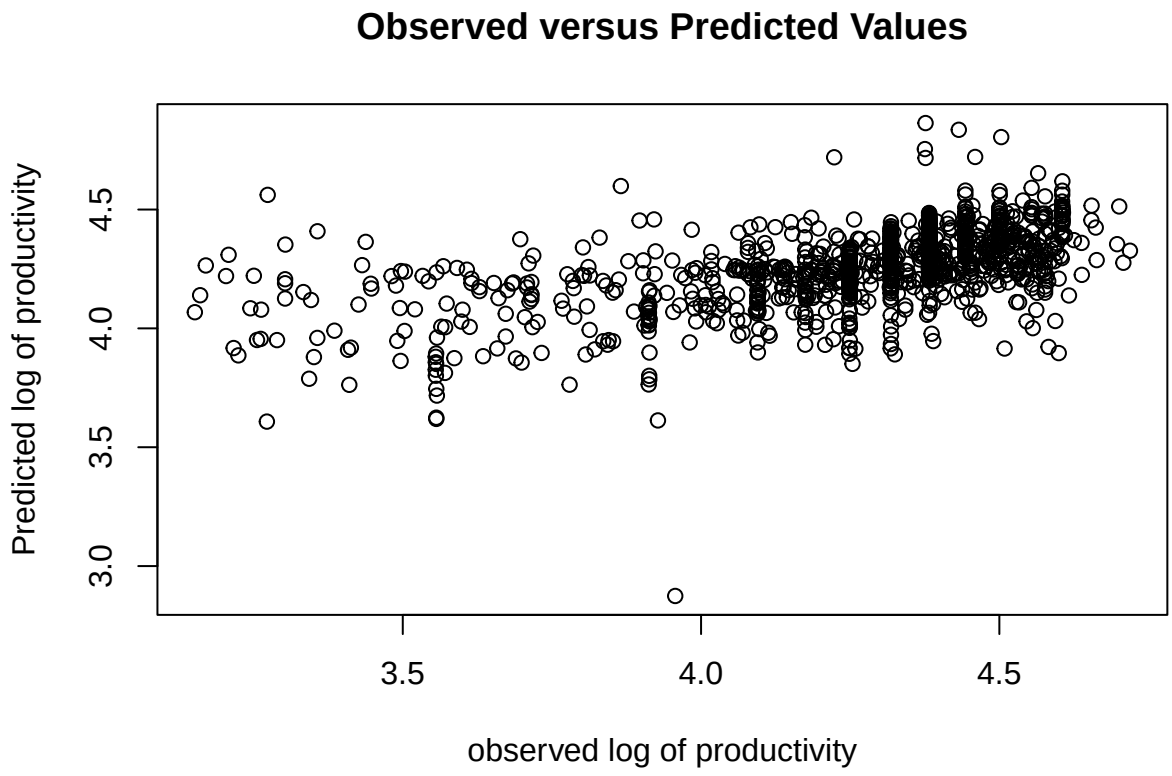
```
4.066351+0+0-0.006396-0.122107
```

```
## [1] 3.937848
```

$E[log\_productivity|(department = Finishing, day = Monday, quarter = 2, team = 8)] = 3.937848+0.178241 \times log(no\_of\_$

vi. Plot the actual log of productivity values versus the predicted log of productivity values. Do you think the model is a good fit? How much variance of the response is explained by the model?

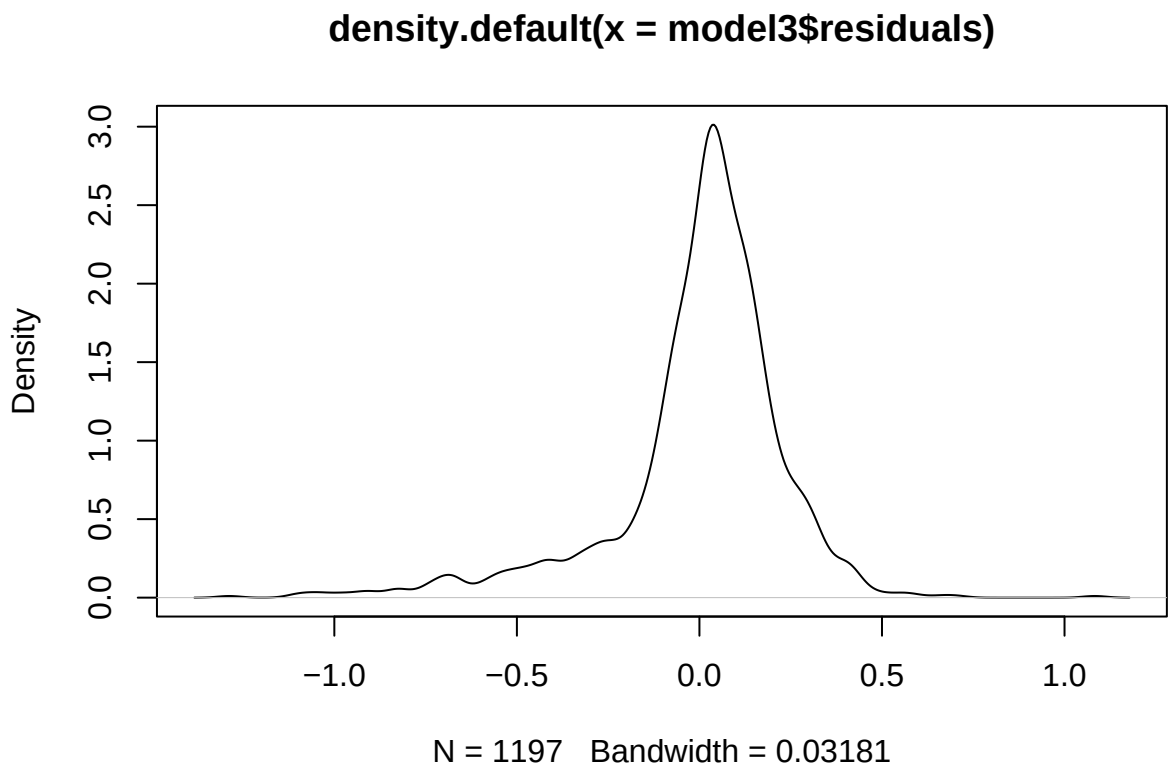
```
pred <- predict(model3, newdata=d)
plot(d$log_productivity,pred,xlab="observed log of productivity",
     ylab="Predicted log of productivity",
     main = "Observed versus Predicted Values")
```



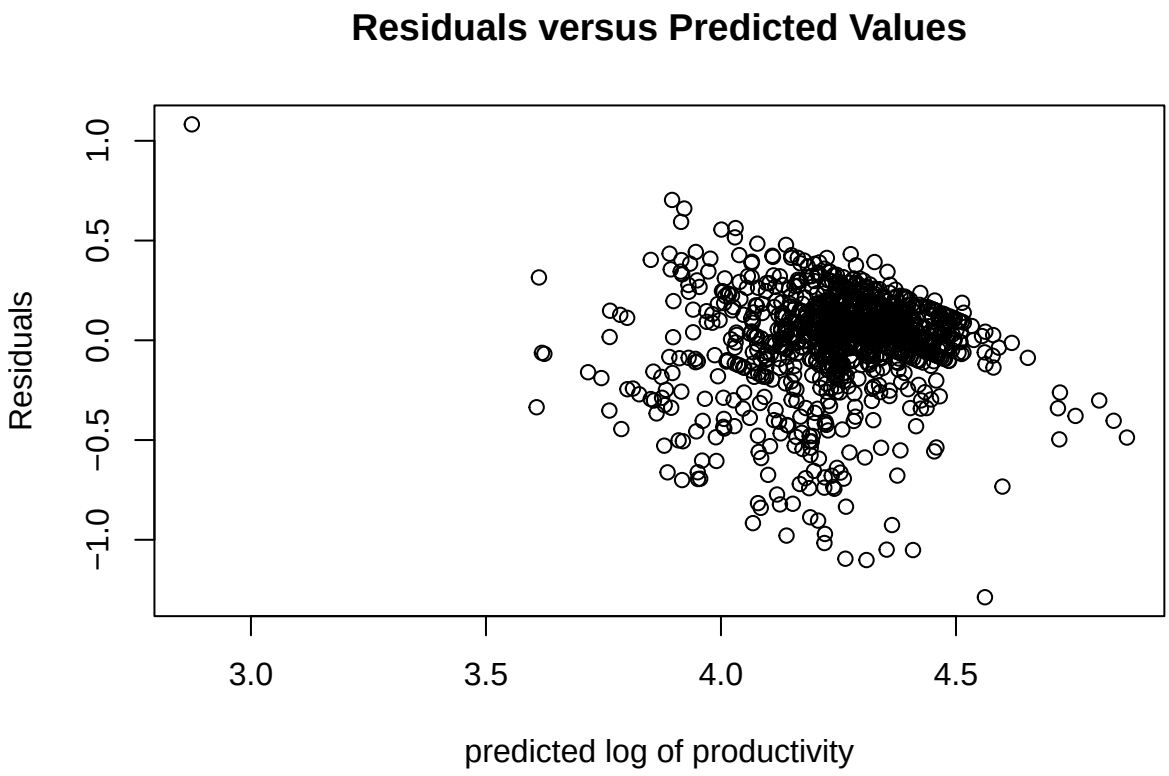
This seems to be a good model. Multiple R-squared is 0.3106 showing that 31.06% of variance of the response is explained by the model.

```
plot(density(model3$residuals))
```

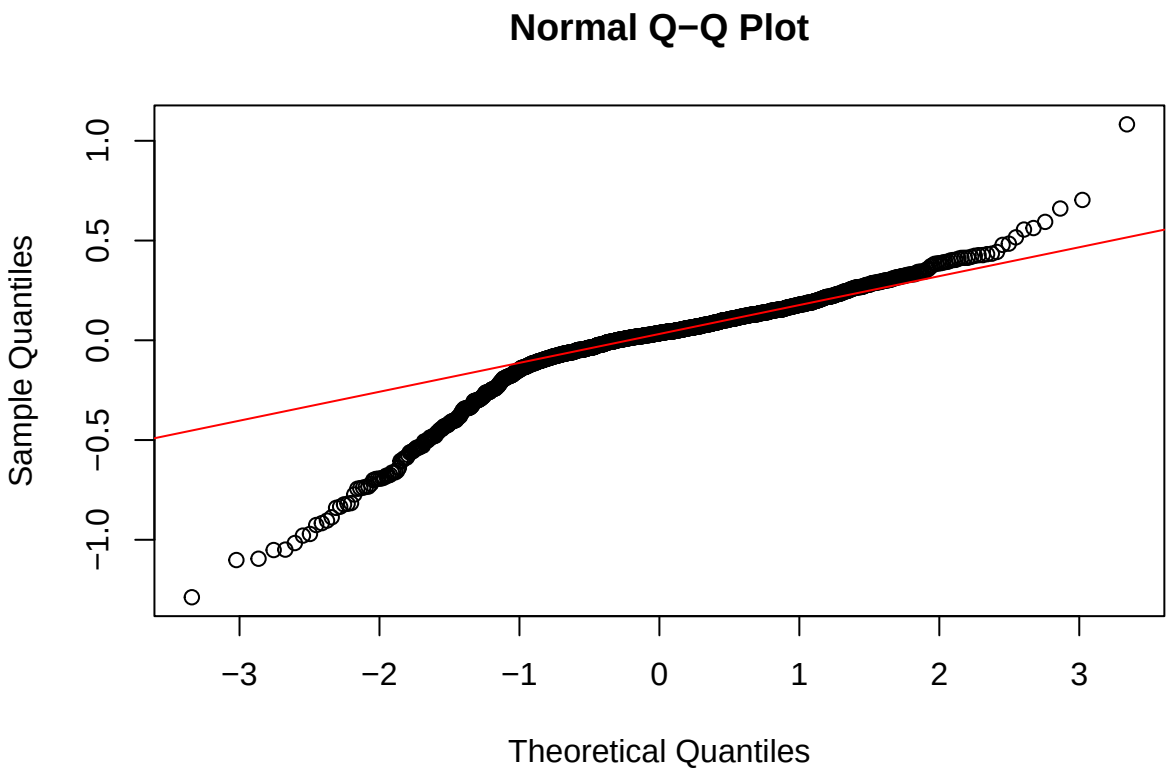
vii. Plot the residuals and the distribution of the residuals. Plot the qqnorm and qqline of the residuals.




```
plot(pred, model3$residuals, xlab="predicted log of productivity",
     ylab="Residuals",
     main = "Residuals versus Predicted Values")
```




```
qqnorm(model3$residuals)
qqline(model3$residuals, col = "red")
```



The residuals are fairly along the diagonal of the qq-plot (quantile-quantile plot). This means that the quantiles of the error distribution are aligned with the quantiles of the reference normal distribution. Therefore, the normality assumption of the error distribution may be valid.

d. Repeat the above (c) for percentage_achievement as the response and all other predictors as above except targeted_productivity.

```
model4 <- lm(percentage_achievment~log(no_of_workers+1)+log(incentive + 1)  + no_of_style_change + quar
summary(model4)
```

```
##
## Call:
## lm(formula = percentage_achievment ~ log(no_of_workers + 1) +
##     log(incentive + 1) + no_of_style_change + quarter + department +
##     day + team, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -80.10  -9.41   1.15   8.49 652.72
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -38.5329    10.4125  -3.701 0.000225 ***
## log(no_of_workers + 1)   19.6702     3.9865   4.934 9.21e-07 ***
## log(incentive + 1)        3.3779     0.7356   4.592 4.86e-06 ***
```

```
## no_of_style_change      -2.9608     2.3817  -1.243 0.214067
## quarterQuarter2          0.9373     2.3610   0.397 0.691431
## quarterQuarter3         -2.7193     2.7090  -1.004 0.315681
## quarterQuarter4         -2.4098     2.6358  -0.914 0.360765
## quarterQuarter5         12.3917     5.0979   2.431 0.015216 *
## departmentsewing       -44.9821     7.0913  -6.343 3.21e-10 ***
## daySaturday              1.1160     3.2073   0.348 0.727935
## daySunday               -0.9604     3.0808  -0.312 0.755289
## dayThursday              3.1363     3.1382   0.999 0.317814
## dayTuesday               2.3167     3.0742   0.754 0.451233
## dayWednesday             1.1366     3.0520   0.372 0.709660
## team2                   -5.1589     4.2157  -1.224 0.221294
## team3                   -1.3939     4.3810  -0.318 0.750418
## team4                   -0.1150     4.2831  -0.027 0.978588
## team5                   -0.8749     4.4242  -0.198 0.843263
## team6                   -5.7170     4.5692  -1.251 0.211111
## team7                   -5.6223     4.3958  -1.279 0.201143
## team8                   -8.1021     4.2483  -1.907 0.056745 .
## team9                  -11.3650     4.2533  -2.672 0.007643 **
## team10                 -10.9563     4.3023  -2.547 0.011005 *
## team11                  -9.4993     4.4678  -2.126 0.033698 *
## team12                  -3.1019     4.4737  -0.693 0.488223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.67 on 1172 degrees of freedom
## Multiple R-squared:  0.08245,    Adjusted R-squared:  0.06366
## F-statistic: 4.388 on 24 and 1172 DF,  p-value: 1.491e-11
```

i. Which of the following variables significantly affect percentage_achievment and which direction? State the level of significance. Log(no_of_workers + 1), log(incentive + 1), department, and team9 are significant at 99.9% level. Quarter5, team10, and team11 are significant at 99% level. Team8 is significant at 95% level. Log(no_of_workers + 1), log(incentive + 1), and quarter5 can cause an increase in percentage_achievment, while the other significant variables can lower percentage_achievment.

ii. On the average how much does percentage_achievment change with one incremental style change. Since the coefficient of no_of_style_change is -2.9608, percentage_achievment would decrease by -2.96% with one incremental style change since there is a log unit.

iii. What is the change in percentage_achievment for quarter 2, 3, 4 and 5 with respect to quarter 1. Which of these changes are statistically significant? Quarter 3, 4 have negative impacts and quarter 2, 5 have positive impacts on percentage_achievment with respect to quarter 1. Quarter 5 is significant at 99% level.

iv. How does the percentage_achievment of sewing department compare with the finishing department? The sewing department has an extremely low percentage achievement since the coefficient of departmentsewing is -44.9821.

v. Write down the regression equation for the following cases:

1. Sewing department for a Sunday of quarter 4 for team 10.

```
-38.5329-0.9604-2.4098 -44.9821 -10.9563
```

```
## [1] -97.8415
```

$E[log\_productivity|(department = Finishing, day = Monday, quarter = 2, team = 8)] = -97.8415+19.6702\times log(no\_of\_$

2. Finishing department for a Wednesday of quarter 1 for team 4.

```
-38.5329+1.1366 -0.1150
```

```
## [1] -37.5113
```

$E[log\_productivity|(department = Finishing, day = Monday, quarter = 2, team = 8)] = -37.5113+19.6702\times log(no\_of\_$

3. Finishing department for a Monday of quarter 2 for team 8.

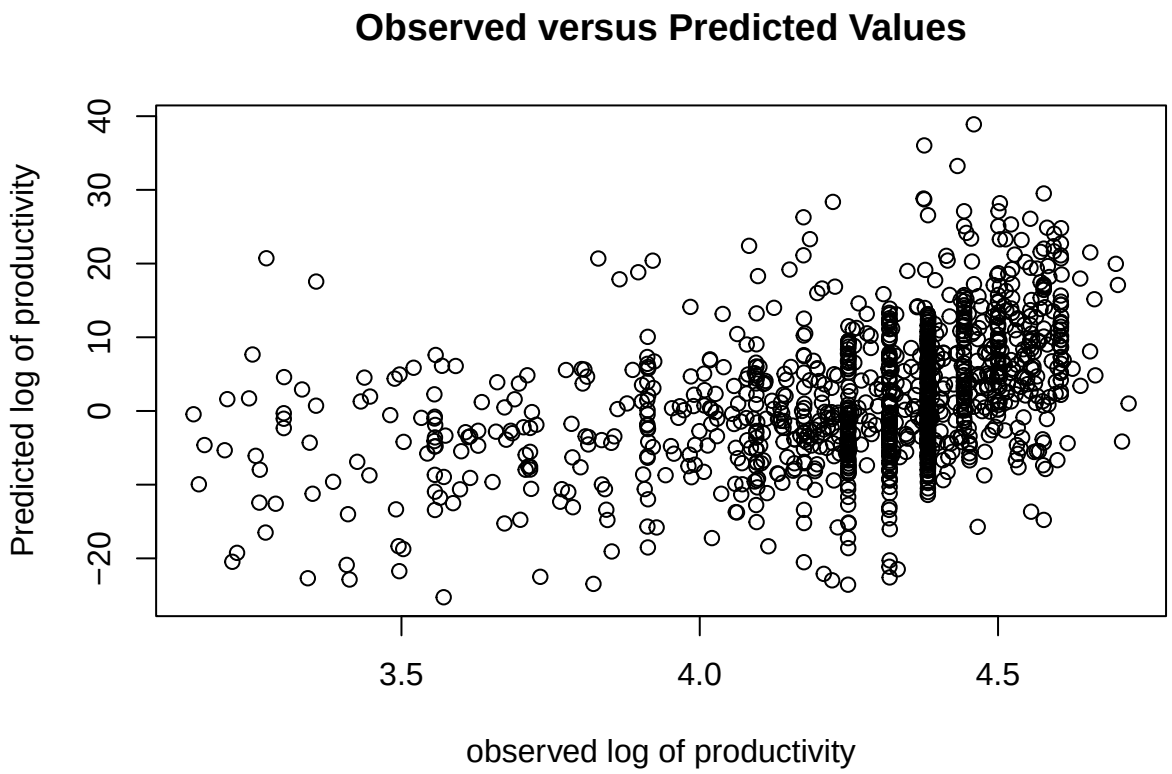
```
-38.5329+ 0.9373 -8.1021
```

```
## [1] -45.6977
```

$E[log\_productivity|(department = Finishing, day = Monday, quarter = 2, team = 8)] = -45.6977+19.6702\times log(no\_of\_$

```
pred2 <- predict(model4, newdata=d)
plot(d$log_productivity,pred2,xlab="observed log of productivity",
     ylab="Predicted log of productivity",
     main = "Observed versus Predicted Values")
```

vi. Plot the actual log of productivity values versus the predicted percentage_achievment values. Do you think the model is a good fit? How much variance of the response is explained by

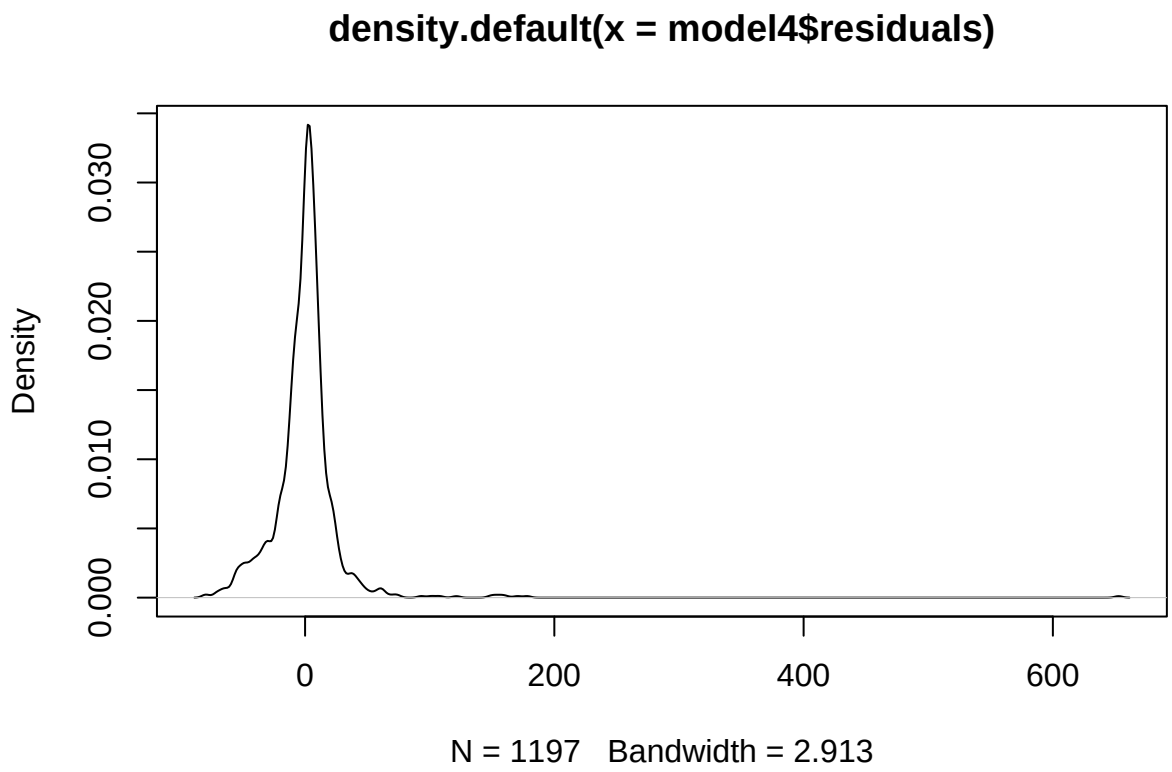


the model?

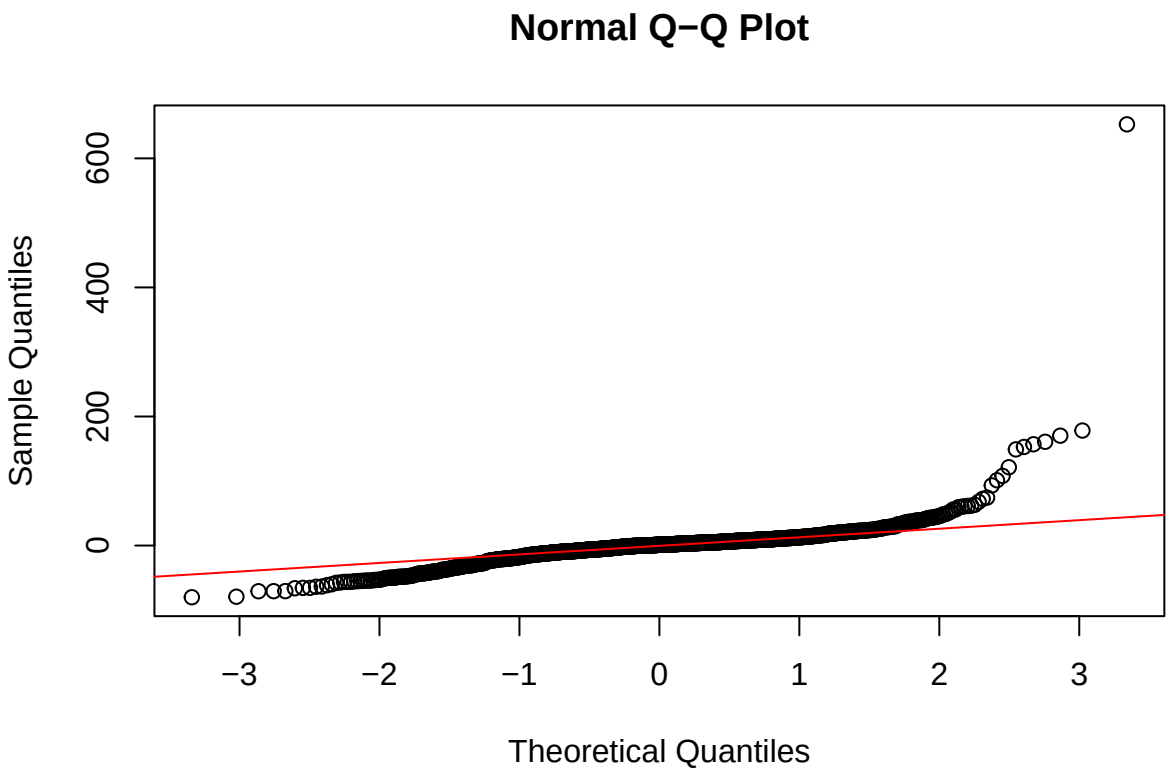
Observation: The model is not performing very well, because the multiple R-squared is 0.08245, which means only 8.245% of variance can be explained by the model.

```
plot(density(model4$residuals))
```

vii. Plot the residuals and the distribution of the residuals. Plot the qqnorm and qqline of the residuals.




```
qqnorm(model4$residuals)
qqline(model4$residuals, col = "red")
```



Observations: The residual points fall roughly along the diagonal line within the qq-plot (quantile-quantile plot) and most of the points roughly follow the qq-line. This means that the quantiles of the error distribution are aligned with the quantiles of the reference normal distribution. Therefore, the normality assumption of the error distribution may be valid.

e. Conduct an ANOVA analysis for question (c) and explain how much (and statistical significance) variance is explained by each variables? Which variable explains the maximum variance?

```
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: log_productivity
##                              Df Sum Sq Mean Sq  F value    Pr(>F)
## log(no_of_workers + 1)        1  0.001  0.0008   0.0139 0.9061712
## log(incentive + 1)            1 10.057 10.0566 170.5631 < 2.2e-16 ***
## log(targeted_productivity)    1 10.212 10.2118 173.1945 < 2.2e-16 ***
## no_of_style_change            1  0.820  0.8200  13.9080 0.0002012 ***
## quarter                       4  0.858  0.2145   3.6375 0.0059300 **
## department                    1  6.304  6.3037 106.9132 < 2.2e-16 ***
## day                           5  0.185  0.0369   0.6267 0.6794706
## team                         11  2.678  0.2434   4.1287 5.418e-06 ***
## Residuals                  1171 69.044  0.0590
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSTotal=0.001+10.057+10.212+0.820+0.858+6.304+0.185+2.678+69.044
SSTotal
```

```
## [1] 100.159
```

```
P_no_of_workers=0.001/SSTotal
P_incentive=10.057/SSTotal
P_targeted_productivity=10.212/SSTotal
P_no_of_style_change=0.820/SSTotal
P_quarter=0.858/SSTotal
P_department = 6.304 /SSTotal
P_day =0.185 /SSTotal
P_team=2.678 /SSTotal
print(c(P_no_of_workers,P_incentive,P_targeted_productivity,P_no_of_style_change,P_quarter,P_department
```

```
## [1] 9.984125e-06 1.004103e-01 1.019579e-01 8.186983e-03 8.566379e-03
## [6] 6.293993e-02 1.847063e-03 2.673749e-02
```

targeted_productivity explains the most.

4. Managerial Insights

Summarize your findings from the above analysis. What can managers of garment manufacturing units learn from your analysis of the data? If a manager is interested in improving the productivity of a garment manufacturing unit, what actions would you suggest (reasonable actions, you cannot ask to stop functioning of a division) to adopt?

Observations

Part2 Part2: 2(b)t-tests performed on quarters 1-5, shows that on average Q5 yields higher prod. 2(c) t-tests performed on department, no_style_change and day shows that only changes of no_style_change is statistically significant to prod (especially the 0-1 change).2(d) and 2(e) shows that incentive is positively correlated with prod compared to #workers. And that incentive and #workers does not correlate very much with %achievement.

model 3(c) is better at explaining var than 3(d) Part 3: targeted_productivity explains the most and the results show that sewing department has lower productivity (department plays role in determining prod). An increase in worker numbers/incentive/target prod can also increase worker prod. Team 7-11 performance (compared to T1 negative) and quarters 4&5 (Q5 positive impact, Q4 negative impact compared to Q1) are significant to prod. Based on 3(a) and 3(b), we can conclude that incentive is statistically significant at predicting prod, but no_workers is not.

####If management want to increase prod then they should increase incentives, and enhance prod in Q5 (or examine reasons why Q5 yields highest prod).