# MSc.IDS - Machine Learning I

Report of Group Work 'Garments Worker Productivity'

Stefan Berchtold, Gianni Pinelli, Stefan Hüttenmoser

11.06.2021

## Contents



### Purpose of the report

The pupose of this report is to show the progress made during the course "Machine Learning I" in the Master's programm "Applied Information and Data Science" at the Hochschule Luzern (HSLU). The goals is to use all the methods taught during the course on a self-chosen dataset. We were free to choose any data set as long as it fulfills following requirements:

-Moderate size(N= [10^3, 10^5]), 10-20 predictors
-Real data
-Contain both continous and categorical variables
-At least one categorical variable must have more than two levels

### Dataset

The dataset contains important attibutes of the garment manufacturing process and the productivity of the employees. It contains following variables:

-date: Date in MM-DD-YYY
-day: Day of the week
-quarter: One-fourth of a year
-no_of_workers: Number of workers in a particular team at a certain time
-team: Number ranging from 1 to 12 for different teams
-no_of_style_change: Number of changes in the style of a particular product
-targeted_productivity: Targeted productivity set by the manager for each team for each day, ranges from 0.07 to 0.8
-smv: Standard Minute Value, it is the allocated time for a task
-wip: Work in progress. Includes the number of unfinished items for products
-over_time: The amount of overtime by each team in minutes
-incentive: The amount of financial incentive, in Bangladesh-Taka (currency of Bangladesh) that enables or motivates a particular course of action
-idle_time: The amount of time when the production was interrupted due to several reasons
-idle_men: The number of workers who were idle due to production interruption
-actual_productivity: The actual % of productivity that was delivered by the workers. It ranges from 0-1

## Data Preaparation

In every Data Science project the first step always has to be the data preparation. This section shows the individual steps

```
garment <-read.csv(file = "~/ML1_project/garments_worker_productivity.csv", stringsAsFactors = FALSE, s
summary(garment)
```

```
##      date              quarter           department            day
##  Length:1197        Length:1197        Length:1197        Length:1197
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       team        targeted_productivity      smv              wip
##  Min.   : 1.000   Min.   :0.0700        Min.   : 2.90   Min.   :    7.0
##  1st Qu.: 3.000   1st Qu.:0.7000        1st Qu.: 3.94   1st Qu.:  774.5
##  Median : 6.000   Median :0.7500        Median :15.26   Median : 1039.0
##  Mean   : 6.427   Mean   :0.7296        Mean   :15.06   Mean   : 1190.5
##  3rd Qu.: 9.000   3rd Qu.:0.8000        3rd Qu.:24.26   3rd Qu.: 1252.5
##  Max.   :12.000   Max.   :0.8000        Max.   :54.56   Max.   :23122.0
##                                                         NA's   :506
##    over_time       incentive         idle_time           idle_men
##  Min.   :    0   Min.   :   0.00   Min.   :  0.0000   Min.   : 0.0000
##  1st Qu.: 1440   1st Qu.:   0.00   1st Qu.:  0.0000   1st Qu.: 0.0000
##  Median : 3960   Median :   0.00   Median :  0.0000   Median : 0.0000
##  Mean   : 4567   Mean   :  38.21   Mean   :  0.7302   Mean   : 0.3693
##  3rd Qu.: 6960   3rd Qu.:  50.00   3rd Qu.:  0.0000   3rd Qu.: 0.0000
##  Max.   :25920   Max.   :3600.00   Max.   :300.0000   Max.   :45.0000
##
##  no_of_style_change no_of_workers   actual_productivity
##  Min.   :0.0000     Min.   : 2.00   Min.   :0.2337
##  1st Qu.:0.0000     1st Qu.: 9.00   1st Qu.:0.6503
##  Median :0.0000     Median :34.00   Median :0.7733
##  Mean   :0.1504     Mean   :34.61   Mean   :0.7351
##  3rd Qu.:0.0000     3rd Qu.:57.00   3rd Qu.:0.8503
##  Max.   :2.0000     Max.   :89.00   Max.   :1.1204
##
```

```
str(garment)
```

```
## 'data.frame':    1197 obs. of  15 variables:
##  $ date                 : chr  "1/1/2015" "1/1/2015" "1/1/2015" "1/1/2015" ...
##  $ quarter              : chr  "Quarter1" "Quarter1" "Quarter1" "Quarter1" ...
##  $ department           : chr  "sweing" "finishing" "sweing" "sweing" ...
##  $ day                  : chr  "Thursday" "Thursday" "Thursday" "Thursday" ...
##  $ team                 : int  8 1 11 12 6 7 2 3 2 1 ...
##  $ targeted_productivity: num  0.8 0.75 0.8 0.8 0.8 0.8 0.75 0.75 0.75 0.75 ...
##  $ smv                  : num  26.16 3.94 11.41 11.41 25.9 ...
##  $ wip                  : int  1108 NA 968 968 1170 984 NA 795 733 681 ...
##  $ over_time            : int  7080 960 3660 3660 1920 6720 960 6900 6000 6900 ...
##  $ incentive            : int  98 0 50 50 50 38 0 45 34 45 ...
##  $ idle_time            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ idle_men             : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ no_of_style_change   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_workers        : num  59 8 30.5 30.5 56 56 8 57.5 55 57.5 ...
##  $ actual_productivity  : num  0.941 0.886 0.801 0.801 0.8 ...
```

The only column which has NA's is wip, it is reasonable to assume that for work in progess a NA indicates that no work is in progress so we change the NA's to 0. To do classification task we take the difference of actual_productivity and targeted_productivity (productivity_difference) and transform it to 1, if actual_productivity is bigger or equal than targeted_productivity (that means that the productivity goals are met or exceeded). If actual_productivity is smaler than targeted_productivity it is transformed to 0 (that means that the productivity goals are not met).The new file is saved as a RDS.

```
garment[is.na(garment)]<- 0
attach(garment)
garment$productivity_difference <- actual_productivity-targeted_productivity
garment$productivity_reached <- ifelse(garment$productivity_difference>=0,1,0)
saveRDS(garment, file = 'garments.rds')
```

In the next step we load the RDS file and delete the column date, since it is not used in the analysis. Since there are still columns which should contain categorical variables but have different data types namely: quarter, department, day and team, we transform those columns to factors (the data type of categorical variables in R)

```
library(dplyr)
df.garment <- readRDS('garments.rds')
df.garment <- select(df.garment,-date)
cols <- c('quarter', 'department', 'day', 'team')
df.garment[cols] <-lapply(df.garment[cols], factor)
```

Now the data is ready to be analyzed we start with a linear model.

## Linear Model

As a first step we fit a linear model, with actual_productivity as dependent variable, and all the remaining variables, besides productivity_difference and productivity_reached because they are based on actual_productivity and might therefore skew the results, as independent varialbes.

```
lm.garment.0 <- lm(actual_productivity ~ .-productivity_difference - productivity_reached , data =df.ga
summary(lm.garment.0)
```

```
##
## Call:
## lm(formula = actual_productivity ~ . - productivity_difference -
##     productivity_reached, data = df.garment)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55763 -0.05958  0.01502  0.08233  0.52152
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.965e-01  4.037e-02   7.343 3.91e-13 ***
## quarterQuarter2        3.220e-03  1.132e-02   0.284 0.776167
## quarterQuarter3       -1.263e-02  1.297e-02  -0.974 0.330226
## quarterQuarter4       -1.074e-02  1.265e-02  -0.849 0.395857
## quarterQuarter5        9.069e-02  2.427e-02   3.737 0.000196 ***
## departmentsweing      -7.742e-02  3.374e-02  -2.295 0.021937 *
## daySaturday            1.350e-02  1.553e-02   0.869 0.384893
```

```
## daySunday               8.636e-04  1.483e-02   0.058 0.953561
## dayThursday            -5.176e-03  1.520e-02  -0.341 0.733507
## dayTuesday              2.011e-02  1.483e-02   1.356 0.175418
## dayWednesday            5.123e-03  1.472e-02   0.348 0.727858
## team2                  -4.383e-02  1.992e-02  -2.200 0.027989 *
## team3                  -7.535e-03  2.073e-02  -0.364 0.716266
## team4                  -2.492e-02  2.024e-02  -1.231 0.218443
## team5                  -5.974e-02  2.107e-02  -2.835 0.004657 **
## team6                  -9.268e-02  2.323e-02  -3.990 7.02e-05 ***
## team7                  -1.027e-01  2.076e-02  -4.948 8.58e-07 ***
## team8                  -9.435e-02  2.014e-02  -4.684 3.15e-06 ***
## team9                  -9.067e-02  2.016e-02  -4.497 7.59e-06 ***
## team10                 -8.994e-02  2.041e-02  -4.407 1.14e-05 ***
## team11                 -1.359e-01  2.209e-02  -6.152 1.05e-09 ***
## team12                 -3.694e-02  2.300e-02  -1.606 0.108546
## targeted_productivity   6.683e-01  4.590e-02  14.560  < 2e-16 ***
## smv                    -7.135e-03  1.020e-03  -6.996 4.43e-12 ***
## wip                     3.809e-06  3.111e-06   1.224 0.221063
## over_time              -4.340e-06  2.030e-06  -2.138 0.032736 *
## incentive               4.481e-05  2.703e-05   1.658 0.097631 .
## idle_time               3.713e-04  4.061e-04   0.914 0.360810
## idle_men               -8.010e-03  1.612e-03  -4.969 7.74e-07 ***
## no_of_style_change     -3.795e-02  1.194e-02  -3.178 0.001521 **
## no_of_workers           5.361e-03  8.701e-04   6.161 9.92e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1453 on 1166 degrees of freedom
## Multiple R-squared:  0.3239, Adjusted R-squared:  0.3065
## F-statistic: 18.62 on 30 and 1166 DF,  p-value: < 2.2e-16
```

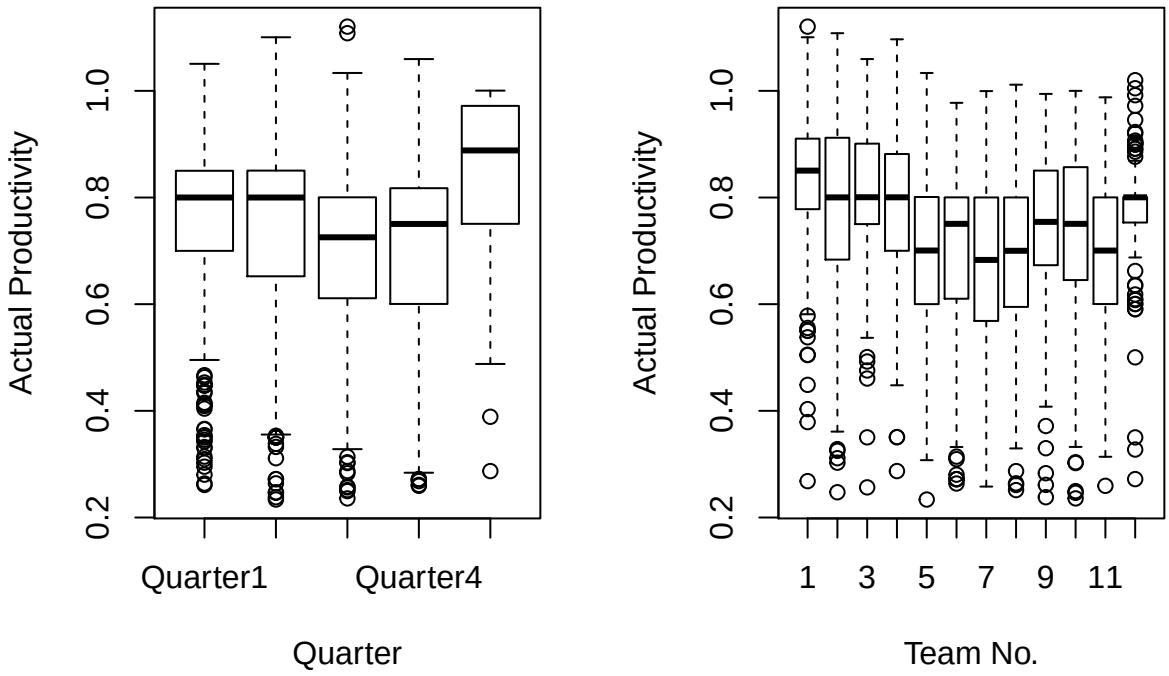
The variables targeted_productivity, idle_men, no_of_worker and most of team seem to have a very strong effect on the response variable. No_of_style_change has a strong effect and department and over_time have a weak effect. The categorical variables like quarter have different signs and team have negative signs. Day seems to have no effect at all and department again has a negativ sign. To check the categorical variables effect on the response variable we test them separately.

```
lm.garmet.cat <- lm(actual_productivity~ department + day + quarter + team, data= df.garment)
drop1(lm.garmet.cat, test = 'F')
```

```
## Single term deletions
##
## Model:
## actual_productivity ~ department + day + quarter + team
##            Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                  32.177 -4284.7
## department  1   0.21734 32.394 -4278.7  7.9365  0.004926 **
## day         5   0.10188 32.279 -4290.9  0.7441  0.590515
## quarter     4   0.73308 32.910 -4265.8  6.6924 2.518e-05 ***
## team       11   3.04153 35.219 -4198.6 10.0969 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variable day seems not to have a effect on actual_productivity, since we assume that it does not interact with any other variable we can drop that variable for the next model. The variables quarter and team seem to be associated with the response variable to better see how, we visualize them.

```r
par(mfrow = c(1, 2))
plot(factor(quarter), actual_productivity, ylab = 'Actual Productivity', xlab = 'Quarter')
plot(factor(team), actual_productivity, ylab = 'Actual Productivity', xlab = 'Team No.')
```




The first plot shows that, the average actual_productivity in quarter 5 was higher than the remaining quarters, and that the productivity seems to drop after the first two quarters. Those are nice insights but this variable will hardly contribute to predict the actual productivity. Furthermore, it will not help the business to take actions to increase the the actual productivity e.g. increase the quarter to 5 and the actual productivity will increase 0.09 (if every other variable stays the same) does not make much sense.
The second plot shows that team number one is, as the name says, number one reagrding actual average productivity. But like the variable quarter it will probably not help to predict the future acttual productivity nor will it help the management to take actions to increase the productivity (again decreasing team number to 1 to achieve higher actual productivity does not make sense).
Those two varialbes have descriptive power but lack predeîctive power, whereas department could have predictive power as well and does make more sense for management action, e.g. shift more volume, faster from sewing to finishing might increase the actual productivitiy.

```r
lm.garment.1 <- lm(actual_productivity ~ .-productivity_difference - productivity_reached -day - team -
summary(lm.garment.1)
```

```
##
## Call:
## lm(formula = actual_productivity ~ . - productivity_difference -
##     productivity_reached - day - team - quarter, data = df.garment)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.54881 -0.06351  0.02230  0.07895  0.50375
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.115e-01  3.534e-02   5.984 2.88e-09 ***
## departmentsweing       -1.058e-01  2.619e-02  -4.038 5.73e-05 ***
## targeted_productivity   6.991e-01  4.595e-02  15.216  < 2e-16 ***
## smv                    -5.873e-03  9.843e-04  -5.966 3.20e-09 ***
## wip                     5.036e-06  3.146e-06   1.601   0.1097
## over_time              -4.338e-06  2.037e-06  -2.129   0.0334 *
## incentive               5.129e-05  2.739e-05   1.872   0.0614 .
## idle_time               4.624e-04  4.174e-04   1.108   0.2681
## idle_men               -9.056e-03  1.648e-03  -5.495 4.78e-08 ***
## no_of_style_change     -4.817e-02  1.165e-02  -4.134 3.81e-05 ***
## no_of_workers           5.422e-03  7.387e-04   7.340 3.95e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1508 on 1186 degrees of freedom
## Multiple R-squared:  0.2594, Adjusted R-squared:  0.2531
## F-statistic: 41.53 on 10 and 1186 DF,  p-value: < 2.2e-16
```

```
RSS.0 <- c(crossprod(lm.garment.0$residuals))
MSE.0 <- RSS.0 / length(lm.garment.0$residuals)
RMSE.0 <- sqrt(MSE.0)
paste('The RMSE of lm.garment.0 is',RMSE.0)
```

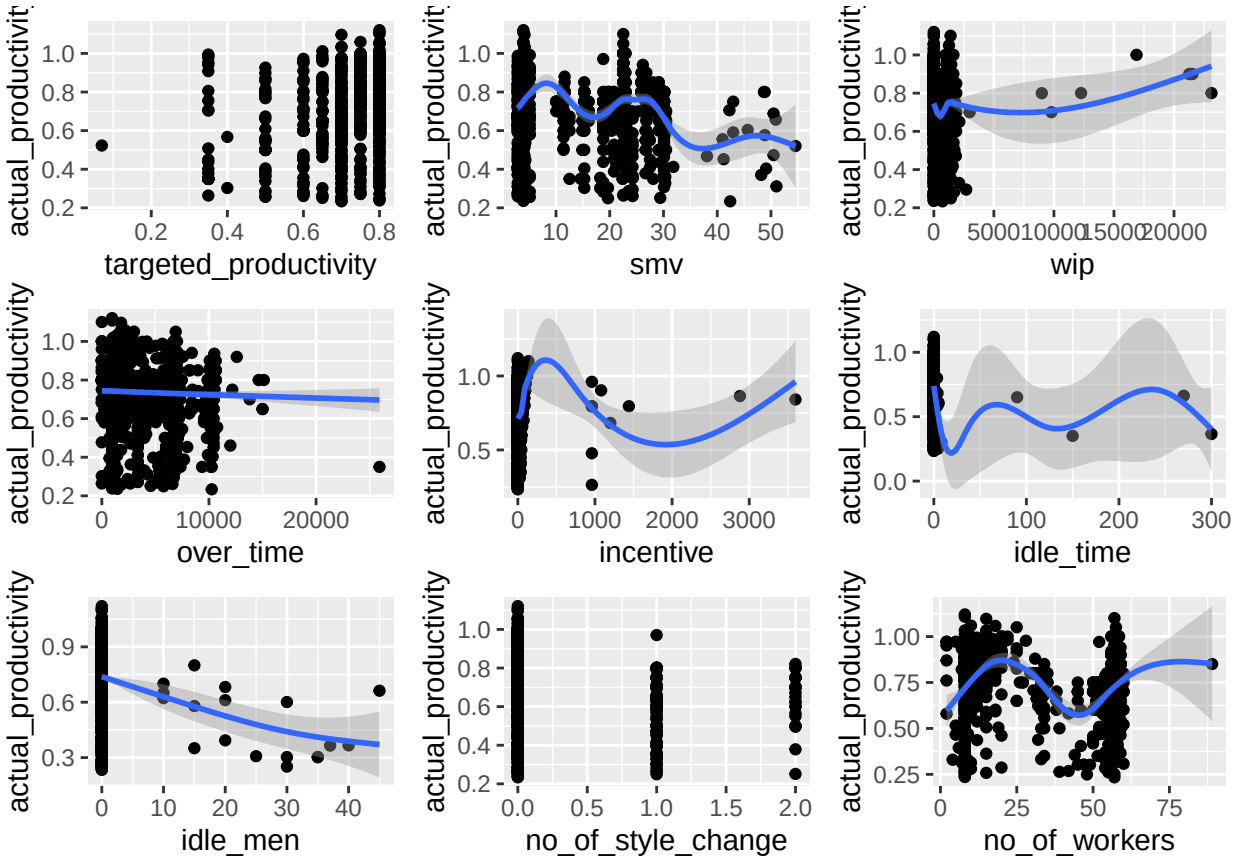
```
## [1] "The RMSE of lm.garment.0 is 0.14341444814118"
```

```
RSS.1 <- c(crossprod(lm.garment.1$residuals))
MSE.1 <- RSS.1 / length(lm.garment.1$residuals)
RMSE.1 <- sqrt(MSE.1)
RMSE.1
```

```
## [1] 0.1501022
```

```
library(ggplot2)
library(gridExtra)

plot1 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = targeted_productivity)) +
geom_point()

plot2 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = smv)) +
geom_point()


plot3 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = wip)) +
geom_point()
```

```r
plot4 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = over_time)) +
geom_point()


plot5 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = incentive)) +
geom_point()


plot6 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = idle_time)) +
geom_point()

plot7 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = idle_men)) +
geom_point()



plot8 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = no_of_style_change)) +
geom_point()

plot9 <- ggplot(data = df.garment,
mapping = aes(y = actual_productivity,
x = no_of_workers)) +
geom_point()


grid.arrange(plot1 + geom_smooth(), plot2 + geom_smooth(),plot3 + geom_smooth(), plot4 +geom_smooth(),pl
```




```
lm.garment.2 <- lm(actual_productivity~department+ targeted_productivity + poly(smv, degree = 3) + wip +

summary(lm.garment.2)
```

```
##
## Call:
## lm(formula = actual_productivity ~ department + targeted_productivity +
##     poly(smv, degree = 3) + wip + over_time + incentive + idle_time +
##     idle_men + no_of_style_change + poly(no_of_workers, degree = 3))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60740 -0.05361  0.01618  0.07503  0.47456
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.852e-01  4.583e-02  10.587  < 2e-16 ***
## departmentsweing              -4.070e-01  5.359e-02  -7.595 6.24e-14 ***
## targeted_productivity          7.122e-01  4.504e-02  15.813  < 2e-16 ***
## poly(smv, degree = 3)1         2.143e+00  8.295e-01   2.584 0.009890 **
## poly(smv, degree = 3)2        -1.985e+00  3.768e-01  -5.267 1.65e-07 ***
## poly(smv, degree = 3)3         7.307e-01  2.604e-01   2.806 0.005092 **
## wip                            4.034e-06  3.081e-06   1.309 0.190718
## over_time                     -6.618e-06  2.017e-06  -3.281 0.001065 **
## incentive                      4.787e-05  2.668e-05   1.794 0.073115 .
## idle_time                      4.902e-04  4.063e-04   1.207 0.227851
```

```
## idle_men                          -9.303e-03  1.605e-03  -5.797 8.65e-09 ***
## no_of_style_change                -4.048e-02  1.170e-02  -3.460 0.000560 ***
## poly(no_of_workers, degree = 3)1   4.570e+00  6.829e-01   6.692 3.39e-11 ***
## poly(no_of_workers, degree = 3)2 -9.502e-01  2.040e-01  -4.657 3.57e-06 ***
## poly(no_of_workers, degree = 3)3   5.958e-01  1.538e-01   3.873 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1467 on 1182 degrees of freedom
## Multiple R-squared:  0.301,  Adjusted R-squared:  0.2927
## F-statistic: 36.36 on 14 and 1182 DF,  p-value: < 2.2e-16
```

```
RSS.2 <- c(crossprod(lm.garment.2$residuals))
MSE.2 <- RSS.2 / length(lm.garment.2$residuals)
RMSE.2 <- sqrt(MSE.2)
RMSE.2
```

```
## [1] 0.1458204
```

### Poisson

```
glm.garment.1 <- glm(incentive ~actual_productivity, family = 'poisson', data = df.garment)
summary(glm.garment.1)
```

```
##
## Call:
## glm(formula = incentive ~ actual_productivity, family = "poisson",
##     data = df.garment)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -12.782   -8.615   -5.045    1.322  156.340
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.01122    0.02514   80.00   <2e-16 ***
## actual_productivity  2.13462    0.03123   68.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 139144  on 1196  degrees of freedom
## Residual deviance: 133962  on 1195  degrees of freedom
## AIC: 137395
##
## Number of Fisher Scoring iterations: 7
```

```
exp(coef(glm.garment.1)['actual_productivity'])
```

```
## actual_productivity
##             8.45387
```

```
glm.garment.2 <- glm(over_time ~actual_productivity, family = 'poisson', data = df.garment)
summary(glm.garment.2)
```

```
##
## Call:
## glm(formula = over_time ~ actual_productivity, family = "poisson",
##     data = df.garment)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -100.721   -53.512    -8.918    32.459   208.978
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           8.590760   0.001809 4748.69   <2e-16 ***
## actual_productivity -0.224218   0.002414  -92.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3073791  on 1196  degrees of freedom
## Residual deviance: 3065253  on 1195  degrees of freedom
## AIC: 3076884
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(glm.garment.2)['actual_productivity'])
```

```
## actual_productivity
##            0.799141
```

### Binominal

```
glm.garment.2 <- glm(productivity_reached ~ department + targeted_productivity + smv + wip + over_time +

summary(glm.garment.2)
```

```
##
## Call:
## glm(formula = productivity_reached ~ department + targeted_productivity +
##     smv + wip + over_time + incentive + idle_time + idle_men +
##     no_of_style_change + no_of_workers, data = df.garment)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0903  -0.5260   0.1396   0.3190   1.0507
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.784e-01  9.710e-02   5.957 3.38e-09 ***
## departmentsweing       2.421e-01  7.196e-02   3.364 0.000792 ***
## targeted_productivity  2.744e-03  1.262e-01   0.022 0.982659
## smv                   -1.687e-02  2.704e-03  -6.240 6.08e-10 ***
## wip                    9.833e-06  8.644e-06   1.138 0.255525
## over_time             -2.961e-06  5.597e-06  -0.529 0.596862
## incentive              1.224e-04  7.526e-05   1.626 0.104204
## idle_time              1.150e-03  1.147e-03   1.003 0.316257
```

```
## idle_men              -2.515e-02  4.528e-03  -5.554 3.44e-08 ***
## no_of_style_change    -5.393e-02  3.201e-02  -1.685 0.092322 .
## no_of_workers          8.195e-03  2.030e-03   4.037 5.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1716425)
##
##     Null deviance: 235.38  on 1196  degrees of freedom
## Residual deviance: 203.57  on 1186  degrees of freedom
## AIC: 1300.4
##
## Number of Fisher Scoring iterations: 2
```