Principal Component Analysis (PCA) and Linear Regression

Authors:

Ife Olalekan EBO	**-**	**62197**
Rajab Mohammed IMAM	**-**	**62198**
Abdullahi Isa AHMED	**-**	**62196**

Submitted to:

Professor Patricia CONDE-CESPEDES

isep
École d'ingénieurs du numérique

Presentation Outlines

- Introduction
- Preliminary Analysis: Descriptive statistics
- Principal Component Analysis (PCA)
- Linear Regression
- Feature selection
- Conclusion
- References

Introduction

- Data is synonymous to breath. Just like the humans breath in every other second, several data are also chunk out in the same manner in every sector of human endeavour.

- It is very difficult to process and understand these data, as they can be too many to comprehend and difficult to analyze their dependencies for decision making.

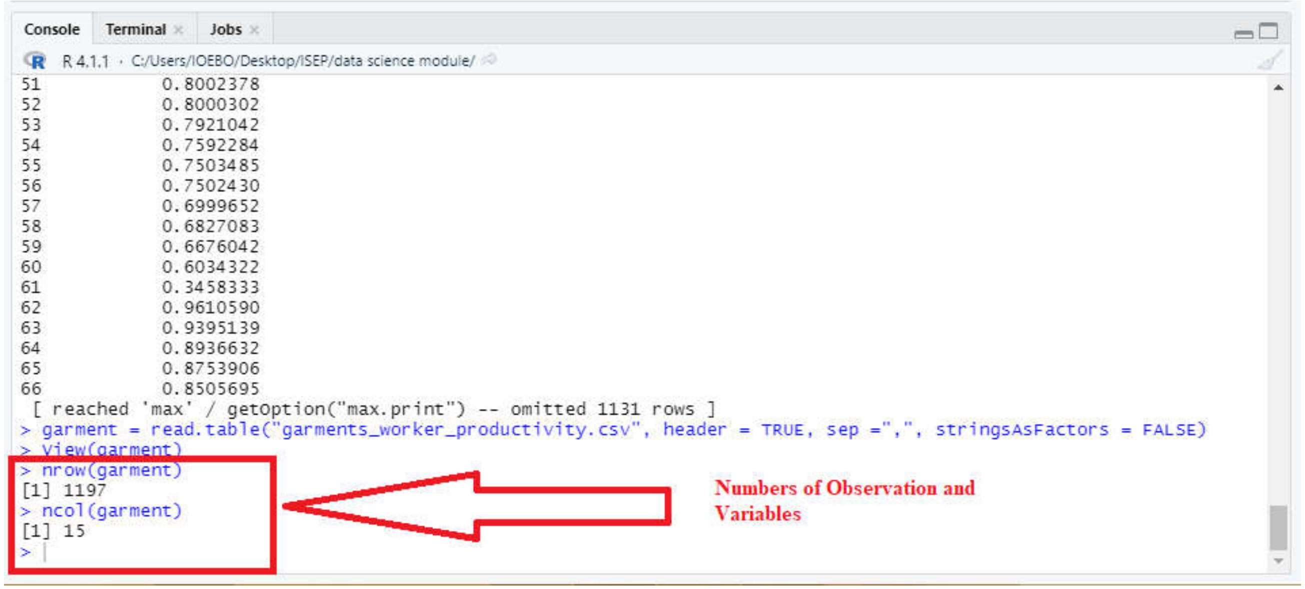
- Thus, the need for a tool such as the Principal Component Analysis (PCA) and Linear Regression helps to visualize and understand these data for optimal decision (present and future) in our day to day activities respectively.

- In this presentation, the PCA and linear regression were used on dataset collected concerning the productivity of workers in the Garment Industry.

Preliminary Analysis: Descriptive statistics (1/5)

From the analysis of the Garment Industry dataset, the following deductions were obtained:

- ❑ There are 1197 Observations and 15 Variables in "*garments worker productivity.csv*" dataset.



```
51            0.8002378
52            0.8000302
53            0.7921042
54            0.7592284
55            0.7503485
56            0.7502430
57            0.6999652
58            0.6827083
59            0.6676042
60            0.6034322
61            0.3458333
62            0.9610590
63            0.9395139
64            0.8936632
65            0.8753906
66            0.8505695
 [ reached 'max' / getOption("max.print") -- omitted 1131 rows ]
> garment = read.table("garments_worker_productivity.csv", header = TRUE, sep =",", stringsAsFactors = FALSE)
> View(garment)
> nrow(garment)
[1] 1197
> ncol(garment)
[1] 15
> 
```

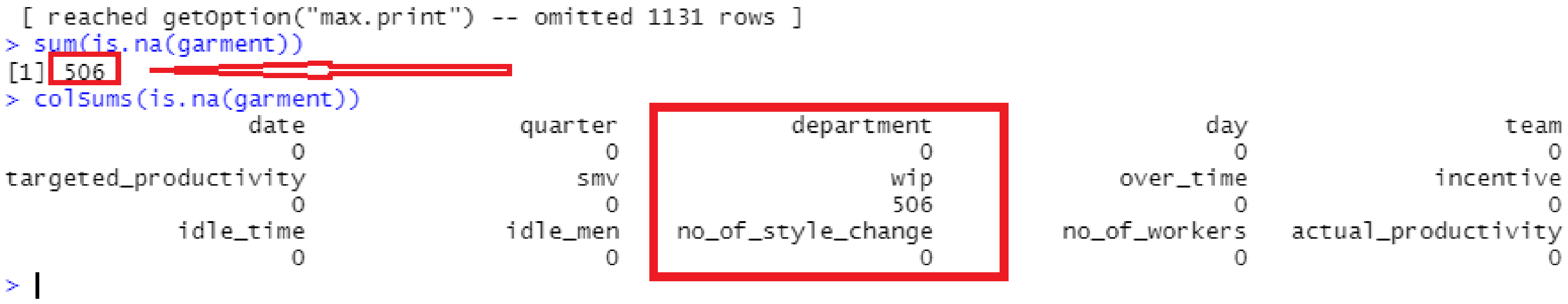
Numbers of Observation and Variables

Figure 1: Observations and variables

Preliminary Analysis: Descriptive statistics (2/5)

Are there missing values?

- **YES**, there are 506 missing values in the dataset under "***wip***" **variable**



```
 [ reached getOption("max.print") -- omitted 1131 rows ]
> sum(is.na(garment))
[1] 506
> colSums(is.na(garment))
                 date               quarter            department                   day                  team
                    0                     0                     0                     0                     0
targeted_productivity                   smv                   wip             over_time             incentive
                    0                     0                   506                     0                     0
            idle_time              idle_men    no_of_style_change         no_of_workers   actual_productivity
                    0                     0                     0                     0                     0
> 
```

Figure 2: Missing values

Preliminary Analysis: Descriptive statistics (3/5)

When dealing with missing data, data scientists can use **Two Primary Methods** to solve error:

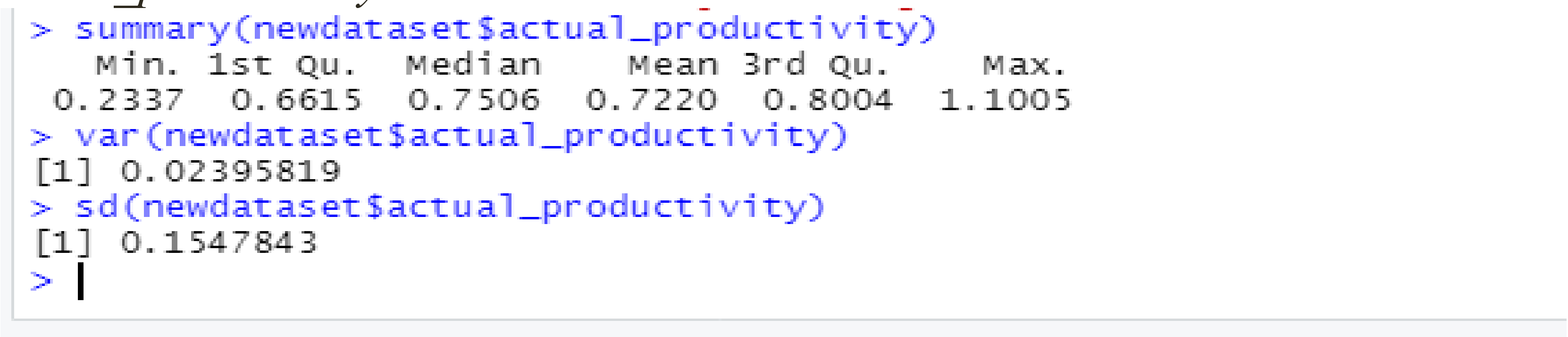
1. The imputation method
2. The removal of data method.

❑ The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

❑ The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Preliminary Analysis: Descriptive statistics (4/5)

A new dataset was generated and all missing values were completely deleted. Below are the descriptive statistics for the target variable "*actual_productivity*".



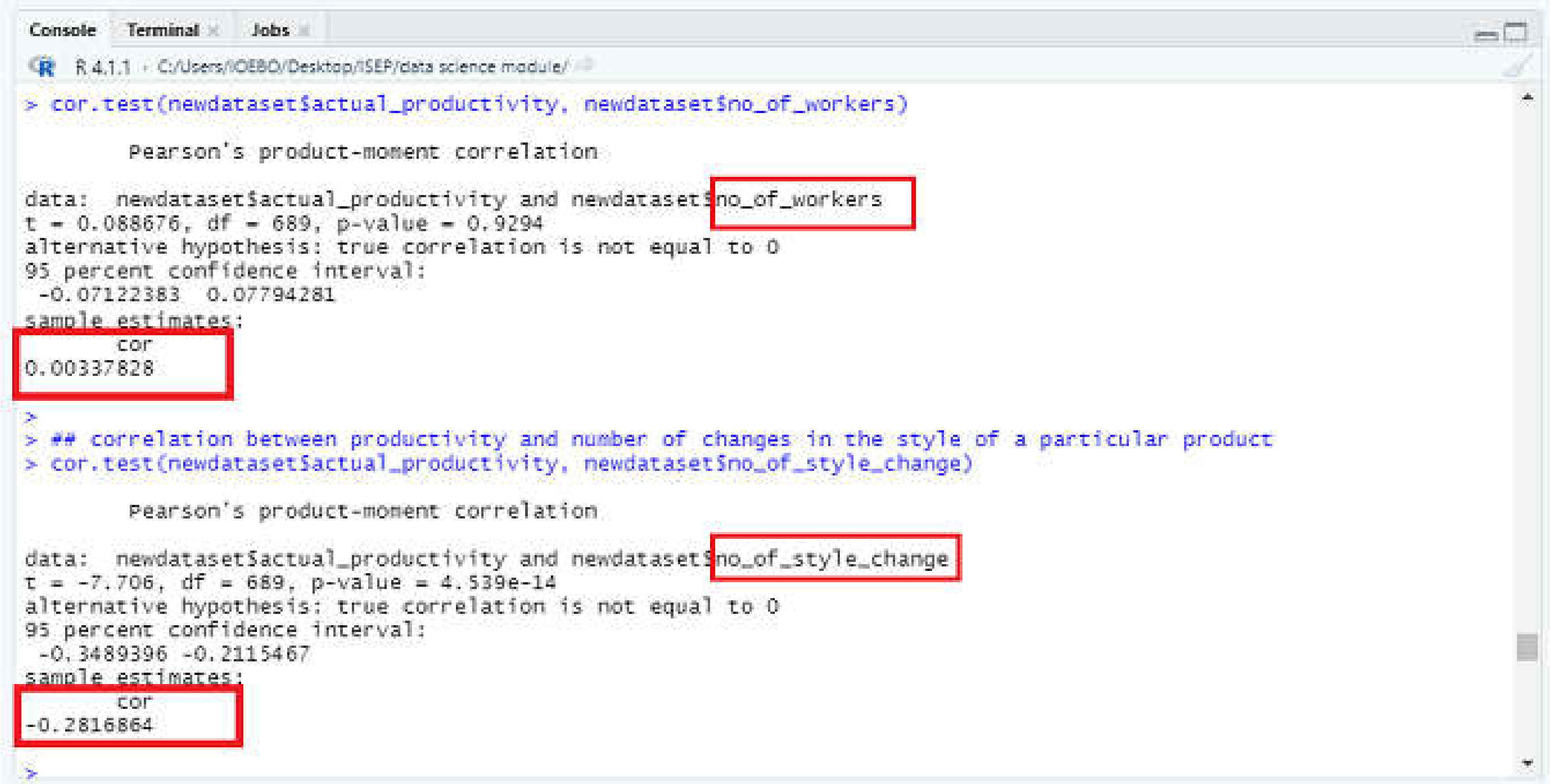
```
> summary(newdataset$actual_productivity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2337  0.6615  0.7506  0.7220  0.8004  1.1005 
> var(newdataset$actual_productivity)
[1] 0.02395819
> sd(newdataset$actual_productivity)
[1] 0.1547843
> 
```

Figure 3: Descriptive statistics of *actual_productivity*

Interpretation: From the output, the average amount of actual productivity is 0.722. Its first quartile (lower 25%) is 0.662, the median (i.e. the lower 50%, second quartile) is 0.751 and third quartile (75%) is 0.800. Its standard deviation is 0.155 which means the data are closely related to the mean, thus a smaller amount of variability (variance is 0.024) in the data, which means it is reliable.

Preliminary Analysis: Descriptive statistics (5/5)

The <span>correlation coefficient</span> between “*actual_productivity*” and other variables are shown below:



```
> cor.test(newdataset$actual_productivity, newdataset$no_of_workers)

	Pearson's product-moment correlation

data:  newdataset$actual_productivity and newdataset$no_of_workers
t = 0.088676, df = 689, p-value = 0.9294
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07122383  0.07794281
sample estimates:
       cor
0.00337828

>
> ## correlation between productivity and number of changes in the style of a particular product
> cor.test(newdataset$actual_productivity, newdataset$no_of_style_change)

	Pearson's product-moment correlation

data:  newdataset$actual_productivity and newdataset$no_of_style_change
t = -7.706, df = 689, p-value = 4.539e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3489396 -0.2115467
sample estimates:
       cor
-0.2816864

>
```

Figure 4. Correlation Coeeficient between “actual_productivity” and other variables

Preliminary Analysis: Descriptive statistics (5/5)

The correlation coefficient between *actual_productivity* and each of the other variables are shown in the table 1 below:

Table 1: Correlation coefficient between *actual_productivity* and other variables

Variables	Correlation Coefficient
no_of_workers	0.00337
no_of_style_change	-0.282
targeted_productivity	0.698
smv	-0.155
wip	0.131
over_time	-0.0168
incentive	0.804
idle_time	-0.113
idle_men	-0.258

Interpretation: The most correlated variables to the *actual_productivity* are ***targeted_productivity*** and ***incentive since their values are closest to ONE (1).***

2.3 Principal Component Analysis (PCA)

It is NOT suitable to include two perfectly correlated variables in the analysis when performing PCA.

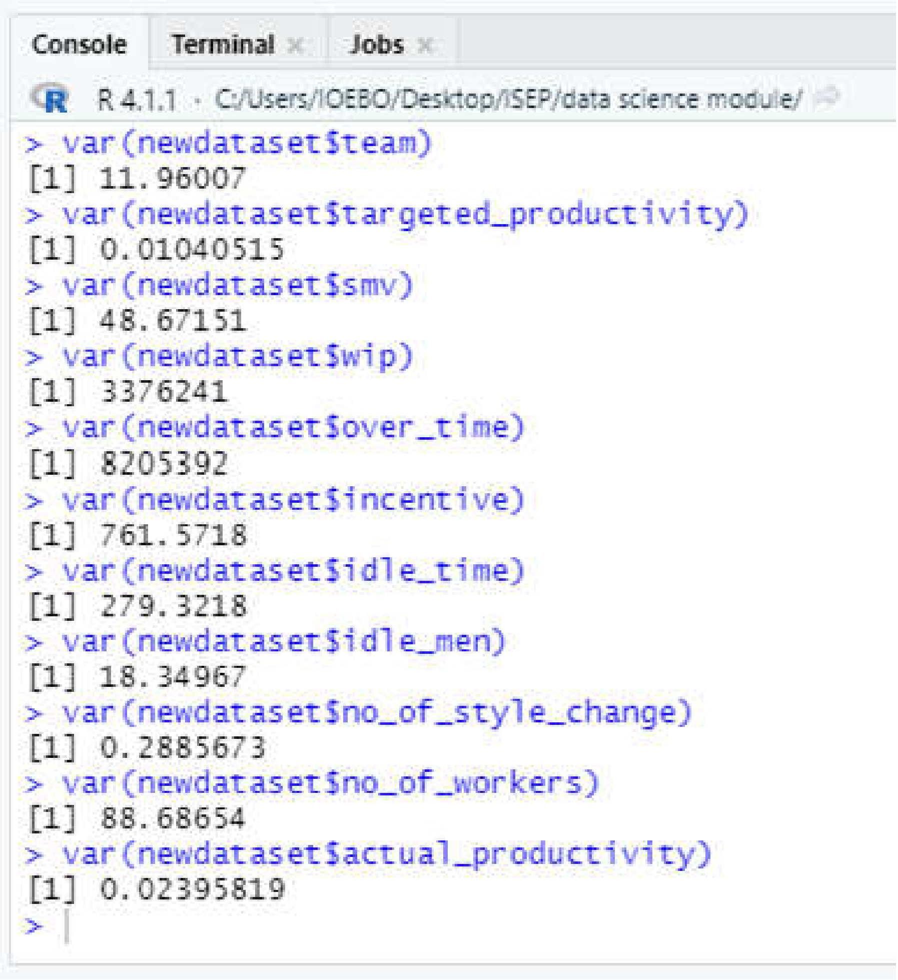
JUSTIFICATION:

- ❑ These correlated variables will cause PCA to over weigh the other variables, thus, overemphasizing their contribution. ***Although, removing them can have a significant impact on the results of PCA.***

Principal Component Analysis (PCA) (1/4)

Interpretation:

Figure 5 shows the variance of each variable. The amount of variability in the "*team*" data is 11.9, "*targeted_productivity*" is "0.010", "*smv*" is 48.7, "*wip*" is 3376241, "*over_time*" is 8205392, "*incentive*" is 761, "*idle_time*" is 279, "*idle_men*" is 18.3, "*no_of_style_change*" is 0.289, "*no_of_workers*" is 88.7 while that of "*actual_productivity*" is 0.024



```
> var(newdataset$team)
[1] 11.96007
> var(newdataset$targeted_productivity)
[1] 0.01040515
> var(newdataset$smv)
[1] 48.67151
> var(newdataset$wip)
[1] 3376241
> var(newdataset$over_time)
[1] 8205392
> var(newdataset$incentive)
[1] 761.5718
> var(newdataset$idle_time)
[1] 279.3218
> var(newdataset$idle_men)
[1] 18.34967
> var(newdataset$no_of_style_change)
[1] 0.2885673
> var(newdataset$no_of_workers)
[1] 88.68654
> var(newdataset$actual_productivity)
[1] 0.02395819
> |
```

Figure 5: Variance of each variables

Principal Component Analysis (PCA) (1/4)

Reasons for Standardization:

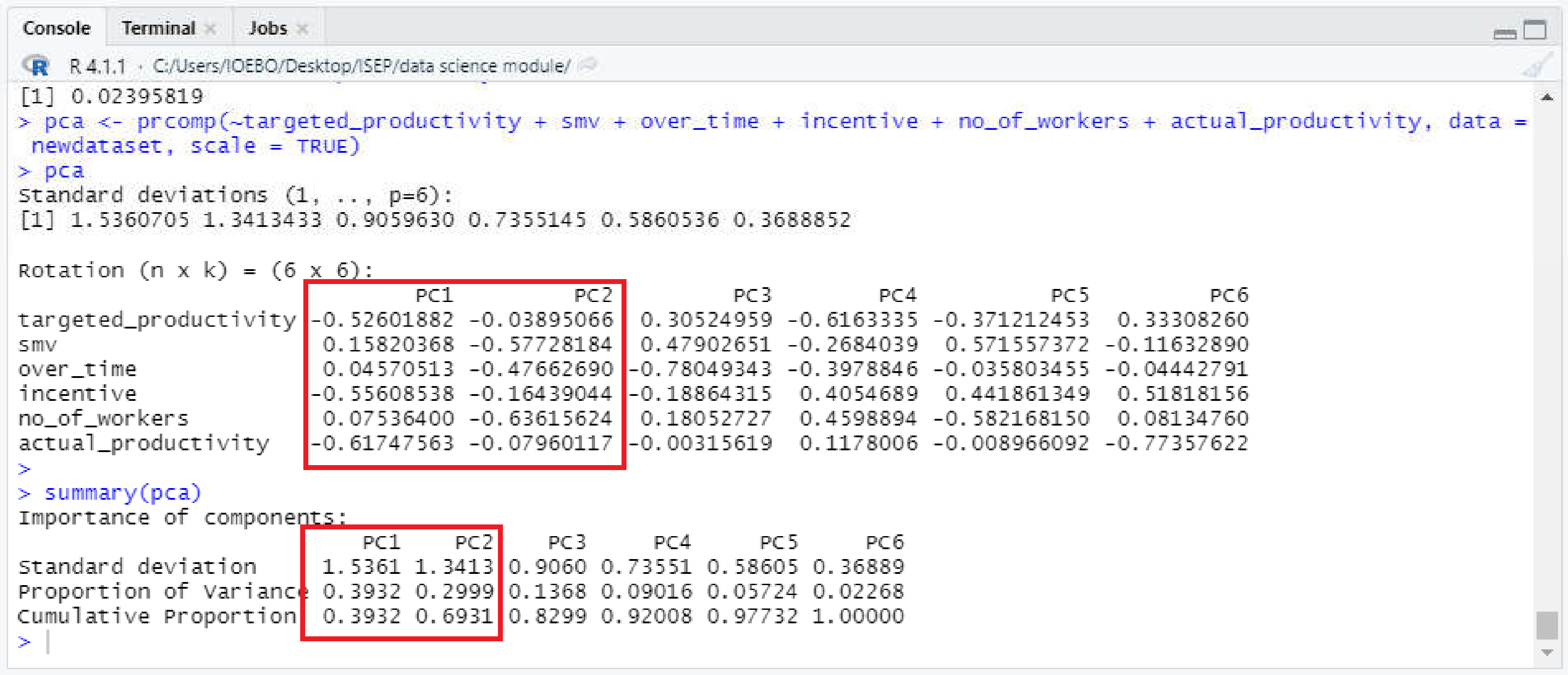
From the output shown in Fig. 5, it is evident that different variables in this dataset have varying variances. Thus, there is an imbalance in weights.

Yes, it is necessary to standardize the dataset before performing the PCA analysis because figure 5 shows an imbalance in the variances of each variables with *over_time* carrying the **largest** value of 8205392 and *targeted_productivity* with the **lowest** value 0.0104.

Figure 5: Variance of each variables

Principal Component Analysis (PCA) (2/4)

Figure 6 shows the PCA result for the six (6) variables using the appropriate function and arguments. The first two PCAs are highlighted with the "*red*" box.



```
[1] 0.02395819
> pca <- prcomp(~targeted_productivity + smv + over_time + incentive + no_of_workers + actual_productivity, data =
 newdataset, scale = TRUE)
> pca
Standard deviations (1, .., p=6):
[1] 1.5360705 1.3413433 0.9059630 0.7355145 0.5860536 0.3688852

Rotation (n x k) = (6 x 6):
                             PC1         PC2         PC3        PC4          PC5         PC6
targeted_productivity -0.52601882 -0.03895066  0.30524959 -0.6163335 -0.371212453  0.33308260
smv                    0.15820368 -0.57728184  0.47902651 -0.2684039  0.571557372 -0.11632890
over_time              0.04570513 -0.47662690 -0.78049343 -0.3978846 -0.035803455 -0.04442791
incentive             -0.55608538 -0.16439044 -0.18864315  0.4054689  0.441861349  0.51818156
no_of_workers          0.07536400 -0.63615624  0.18052727  0.4598894 -0.582168150  0.08134760
actual_productivity   -0.61747563 -0.07960117 -0.00315619  0.1178006 -0.008966092 -0.77357622
>
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.5361 1.3413 0.9060 0.73551 0.58605 0.36889
Proportion of Variance 0.3932 0.2999 0.1368 0.09016 0.05724 0.02268
Cumulative Proportion  0.3932 0.6931 0.8299 0.92008 0.97732 1.00000
> 
```

Figure 6: PCA values

Principal Component Analysis (PCA) (2/4)

Interpretation:

- The first PCA, PC1 (as shown in Fig. 6) accounts for the largest possible amount of variation that may be present in the data. It shows approximately equal weight on targeted_productivity, incentive and actual_productivity with much less weight on smv, overtime and no_of_workers. Hence this component roughly measures the major factor that contributes to workers productivity.

- The second PCA, PC2 accounts for as much of the remaining variation as possible, with the constraint that the correlation between the first and second component is 0. It places most of its weight on smv, overtime and no_of_workers and much less weight on the other three features.

Principal Component Analysis (PCA) (3/4)

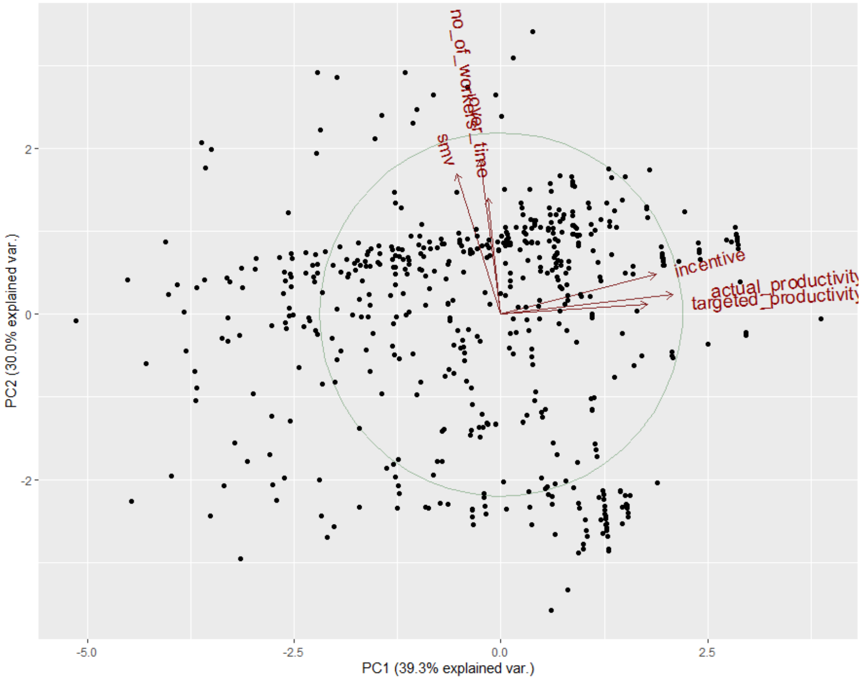


Figure 7: Biplot with correlation circle of PCA scores and Loading Vectors

Interpretation:

- It shows that the variables whose unit vectors are close to each other are said to be positively correlated (i.e. *targeted_prod., incentives and actual_prod.,*), meaning that their influence on the positioning of individuals is similar.
- The same is obtainable with the other 3 variables.
- However, there is a far distant gap between each group of these 3 correlated variables, thus, they are uncorrelated.

Principal Component Analysis (PCA) (4/4)

Table 2: Percentage Variance Explained (PVE) by each components

Components	Percentage (%)
PC1	39
PC2	29
PC3	13
PC4	0.9
PC5	0.5
PC6	0.2

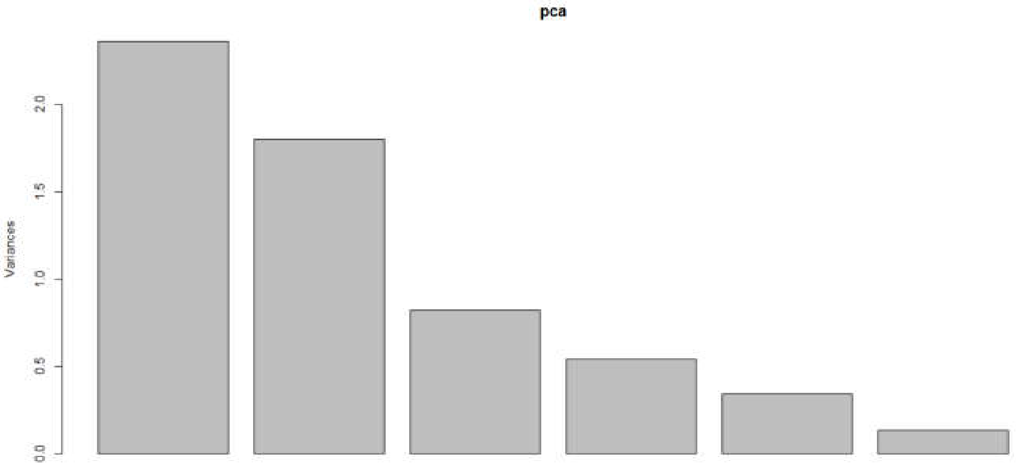


Figure 8a. Percentage Variance Explained (PVE) by each components

Principal Component Analysis (PCA) (4/4)

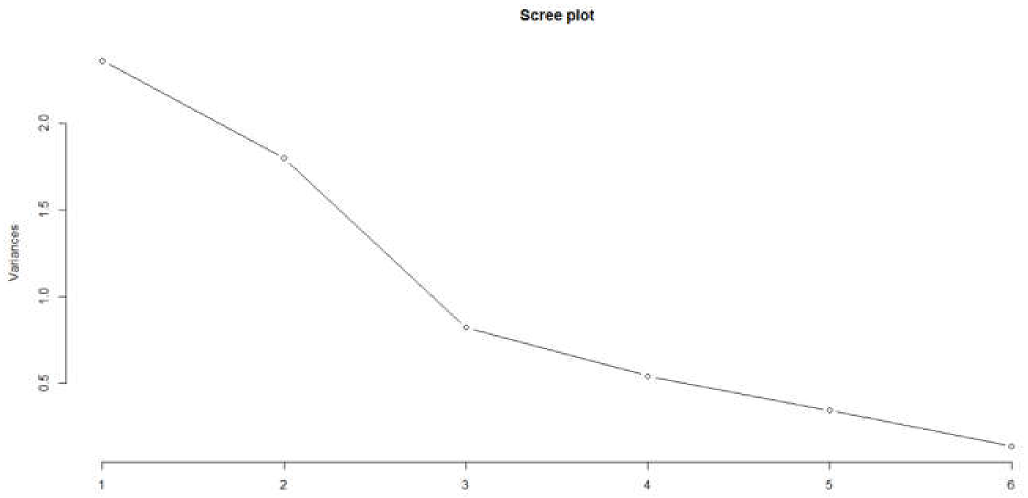


Figure 8b. Scree plot of PVE for each component

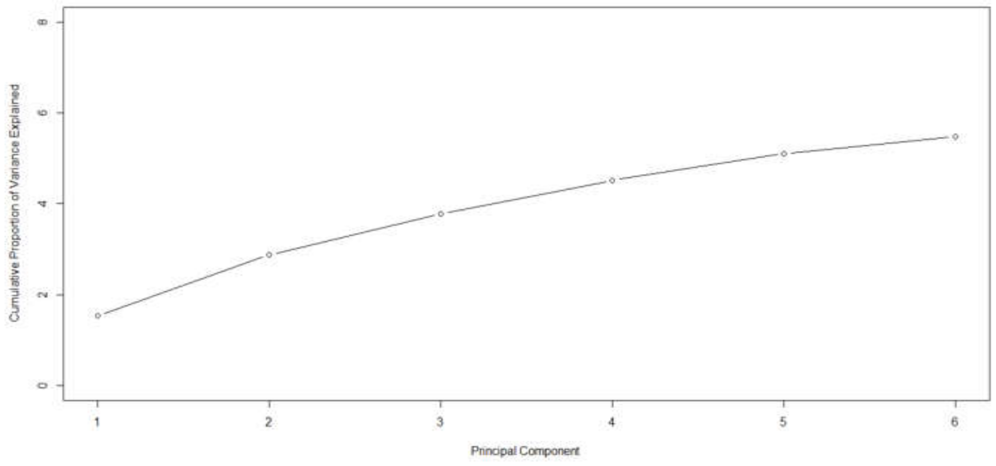


Figure 8c. Cumulative PVE for the components

Principal Component Analysis (PCA) (4/4)

How many components would you keep?

The first three components will be kept, i.e PC1, PC2 and PC3

Why?

The PCAs represents the maximum amount of variance in the data, which means the lines that captures most information in the data.

Linear Regression

Theoretical Question

- **Relationship between r and R2**

The correlation coefficient r tells us the relationship between both variables. Coefficient of determination, R2(X, Y) is the square of the correlation coefficient r (X, Y), which is used to describe the amount of variation in Y explained by X.

- **What is the range of values that can be taken by R2?**

❑ *R2* values ranges from 0 to 1 (0 to 100%)

How about r?

❑ *r* values ranges from -1 to 1

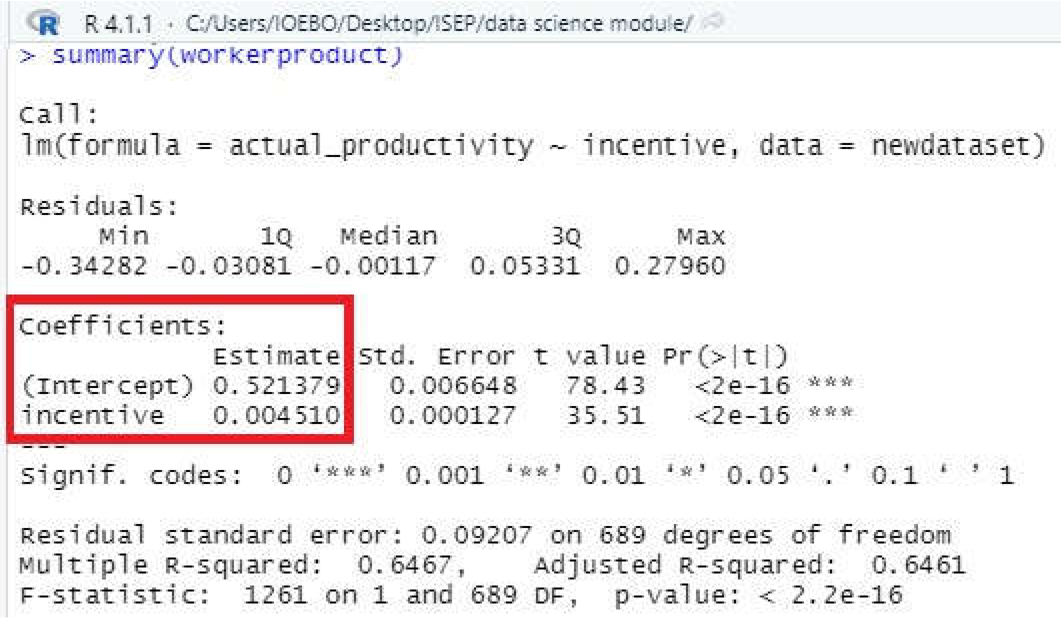
Linear Regression (1/4)

(Practical Application)

What are the coefficient estimates ?

The coefficient estimates, Bo and B1 are 0.521 and 0.00451 respectively as shown in the Fig. 9.The regression equation is:

Productivity = 0.521 + 0.00451incentive



```
R 4.1.1 · C:/Users/IOEBO/Desktop/ISEP/data science module/
> summary(workerproduct)

Call:
lm(formula = actual_productivity ~ incentive, data = newdataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.34282 -0.03081 -0.00117  0.05331  0.27960 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.521379   0.006648   78.43   <2e-16 ***
incentive   0.004510   0.000127   35.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09207 on 689 degrees of freedom
Multiple R-squared:  0.6467,	Adjusted R-squared:  0.6461 
F-statistic:  1261 on 1 and 689 DF,  p-value: < 2.2e-16
```

Figure 9: Coefficient estimate

Interpretation of the coefficient estimate, B1

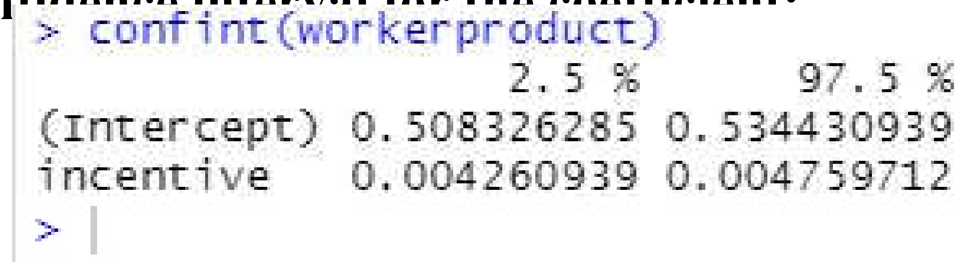
Coefficient of estimate, B1 (0.00451) represents the incentive coefficient in the regression equation. This coefficient represents the mean increase of productivity for every additional oe increase in the incentive. If the incentive increases by 1, the average productivity increases by 0.00451

Linear Regression (2/4)

General expression for CI of (1-∝) for Bi:

CI = [Bi{est} – 1.96 x SE(Bi{est}), Bi{est} + 1.96 x SE(Bi{est})]

95% confidence interval for the coefficient:



```
> confint(workerproduct)
                  2.5 %      97.5 %
(Intercept) 0.508326285 0.534430939
incentive   0.004260939 0.004759712
> |
```

Fig. 10: 95% CI for coefficient

Interpretation:

The Confidence Interval (CI) is (0.00426, 0.00476). This interval can be seen as the set of null hypotheses for which a 5% two-sided hypothesis does not reject.

Linear Regression (3/4)

- From Fig. 9, which shows the model summary, the p-value <2e-16, which is approximately 0 and less than 0.05, thus, there is a linear relationship between productivity and financial incentive.

- This relationship is statistically significant as B1 is significantly non zero.

What is the value of R2?
The value of R2 is 0.6467.

This means that the variable "financial incentive" can explain up to 64.6% of the variations in the variable "actual productivity". The remaining 35.4% of variation can be explained by other variables which we have not included in this model.

Yes, the model is suitable to predict productivity.

Feature Selection (1/5)

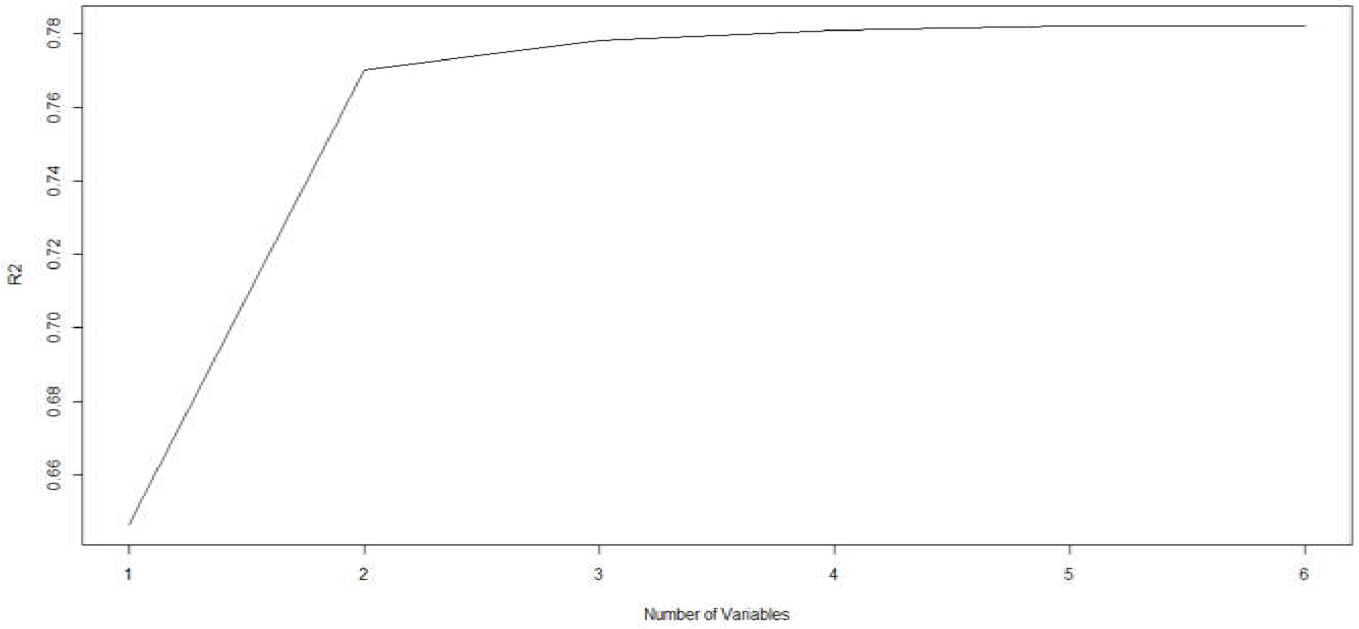
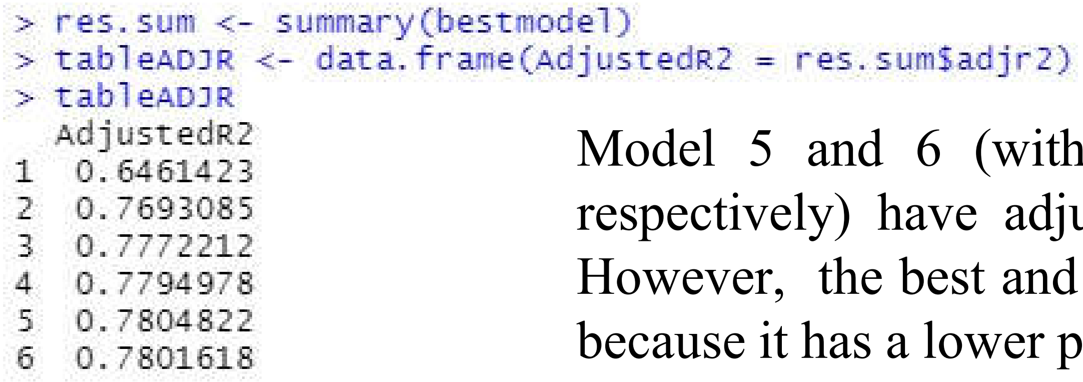


Fig. 11a. R-squared vs number of features



```
> res.sum <- summary(bestmodel)
> tableADJR <- data.frame(AdjustedR2 = res.sum$adjr2)
> tableADJR
  AdjustedR2
1  0.6461423
2  0.7693085
3  0.7772212
4  0.7794978
5  0.7804822
6  0.7801618
```

Model 5 and 6 (with five and six variables respectively) have adjusted R-square of 0.780. However, the best and more suitable model is 5 because it has a lower process capability (CP),

Feature Selection (2/5)

Features Kept:

- Five features were kept. They are: *target_productivity, smv, over_time, incentive, and number of workers.*

Why Adjuted R-Squared and NOT R-squared

- R-square increases as we add more and more predictors to our model, even if it is by chance, hence, the model will appears to have more explanatory power than the model with lesser number of predictors.

- Meanwhile, Adjusted R-squared will only increase if the added predictors improves the model, rather than just increase because more predictors were added. Thus, it represents the proportion of variation, in the outcome, that are explained by the variation in predictors values. the higher the adjusted R2, the better the model.

Feature Selection (3/5)

```
Coefficients:
                  Estimate      Std. Error      t value      Pr(>|t|)
(Intercept)   1.384e-01     2.763e-02       5.007        7.04e-07 ***
targeted_prod  6.216e-01    3.180e-02       19.550       < 2e-16 ***
smv             -2.757e-03  4.987e-04       -5.528       4.60e-08 ***
over_time      -2.108e-06   1.044e-06       -2.019       0.04388 *
incentive      3.319e-03    1.196e-04       27.764        < 2e-16 ***
no_of_workers 1.216e-03     3.787e-04       3.210        0.00139 ** ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.07252 on 685 degrees of freedom
Multiple R-squared:  0.7821,    Adjusted R-squared:  0.7805
F-statistic: 491.7 on 5 and 685 DF,  p-value: < 2.2e-16
```

Interpretation:

- Average *actual_productivity* is 0.138.
- 1 unit change in *targeted_productivity* increases *actual_productivity* by 0.622 given that other variables are constant.
- 1 unit change in *smv* decreases *actual_productivity* by 0.003 given that other variables are constant.
- 1 unit change in *over_time* decreases *actual_productivity* by -0.000002 given that other variables are constant.

Feature Selection (5/5)

- 1 unit change in *incentive* increases *actual_productivity* by 0.003 given that other variables are constant.
- 1 unit change in *no_of_workers* increases *actual_productivity* by 0.0012 given that other variables are constant.

The value of the coefficient of determination R2 is 0.7821

- From the summary of the coefficient estimates table given, the zero slope hypothesis test result is shown.
- Thus, since the p-values of the predictors are less than 0.05, it can be concluded that the effect of all predictors on actual_productivity included is statistically significant.

Conclusion

In this presentation, we examined the dataset collected concerning the productivity of workers in the Garment Industry. We performed descriptive statistics on the dataset, analyse the data and we eliminate the missing values from the dataset to generate a new dataset.

The correlation coefficient between *actual_productivity* and each of the other variables led us to standardization of data due to imbalance discovered and we performed PCA. The first two PCAs were analysed. Finally, we performed linear regression and feature selection on the dataset to find relationships, confidence intervals and deduced that the model is suitable to predict productivity.

References

Centellegher, S. (2020, January 27). *How to compute PCA loadings and the loading matrix with scikit-learn*. Simone Centellegher, PhD - Data Scientist and Researcher. Retrieved January 2022, from https://scentellegher.github.io/machine-learning/2020/01/27/pca-loadings-sklearn.html

Gables, C. (2020, September 22). *How to Interpret correlation coefficient (r)?* STATS-U. Retrieved January 20, 2022, from https://sites.education.miami.edu/statsu/2020/09/22/how-to-interpret-correlation-coefficient-r/

Ghassany, M. (2021, December 1). *E Model Selection | Machine Learning*. Engineering School de Vince Paris. Retrieved January 14, 2022, from https://www.mghassany.com/MLcourse/model-selection.html

Kabacoff, R. I. (2017). *Quick-R: Descriptives*. Quick-R. Retrieved January 2022, from https://www.statmethods.net/stats/descriptives.html

Mahbobi, M. (2015, December 7). *Chapter 8. Regression Basics – Introductory Business Statistics with Interactive Spreadsheets – 1st Canadian Edition*. Pressbooks. Retrieved January 20, 2022, from https://opentextbc.ca/introductorybusinessstatistics/chapter/regression-basics-2/

R linear regression test hypothesis for zero slope. (2013, March 14). Cross Validated. Retrieved January 20, 2022, from https://stats.stackexchange.com/questions/52256/r-linear-regression-test-hypothesis-for-zero-slope

Imran, A. A., Amin, M. N., Islam Rifat, M. R., and Mehreen, S. (2019). "Deep Neural Network Approach for Predicting the Productivity of Garment Employees". In 6th International Conference on Control, Decision and Information Technologies (CoDIT)

Rahim, M. S., Imran, A. A., and Ahmed, T. (2021) : "Mining the Productivity Data of Garment Industry". In International Journal of Business Intelligence and Data Mining, 1(1).

Thank You…!!!

Q & A