Predicting Software Project Delivery

ISQA 8340 - Applied Regression

Adam McQuistan

Contents

Introduction

This breif communication assesses the use of linear models to predict the productivity of software development teams in the early 2000s. The data set under study is provided in the appendix of Kutner, Nachtsheim, and Neter's book Applied Linear Regression Models and describes the productivity of software development teams for a consulting company over the years of 2001 and 2002. The variables collecticed in the study are listed in table 1. Count is the number of website produced by team per quarter over the years 2001 and 2002. Backlog is the number of projects to be completed at the start of each quarter. Team is a categorical variable used to represent the 13 different teams under study. Team experience is the number of months the team had been working together by the end of the quarter. Process is a categorical variable representing two different software development methods that were used over the two years. It is important to note that in the second quarter of year 2002 a new software development method was enacted company wide. Year is a categorical variable for the two years of the study.

Variable	Type	Description
count	ratio	count of website produced in a quarter
backlog	ratio	count of projects in backlog
team	categorical	team label
team_exp	ratio	number of month team has worked together
processA	categorical	1 for process A, 0 for process B
year2001	categorical	1 for 2001 and 0 for 2002

Overview

To predict the productivity a statistical method known as regression will be used and implemented in the R statistical programming language. Linear regression is an analytical technique used to model relationships between one or more input (dependent) variables and a continuous outcome variable with the key assumption that the relationship between the dependent variables and the outcome variables are linear [1]. It is common to use transformations on the outcome or dependent variables to acheive linearity [2]. The resulting linear regression model is a probabilistic one that accounts for randomness and factors not included in the building of the model [1]. Therefore, the model is used to find the expected value of the outcome variable based off the input variables and comes with some level of uncertainty.

Regression is very common and powerful statistical tool for learning interesting things about a particular data set in a way that lends to simple interpretation of the end result. However, it is important that your model does not violate the fundamental assumptions described previously in order to build a reliable or robust regression model. If fact, if one is not careful in their approach you can easily build a misleading model.

Model Description

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \varepsilon$$

- y is the outcome variable
- x_j are the input variables for j=1,2,...,p-1
- β_0 is the value of y where each x_j equals zero
- β_j is the change in y based on one unit change in x_j for j=1,2,...,p-1

- $\varepsilon \sim N(0, \sigma^2)$ is a random error term that represents the difference in the linear model and the observed value of y. The assumption is that the mean of the error term is zero and each ε is independent of each other and normally distributed. The assumption of normal distribution in the error term (residuals) allows for hypothesis testing and confidence interval estimation [1].

Data Analysis

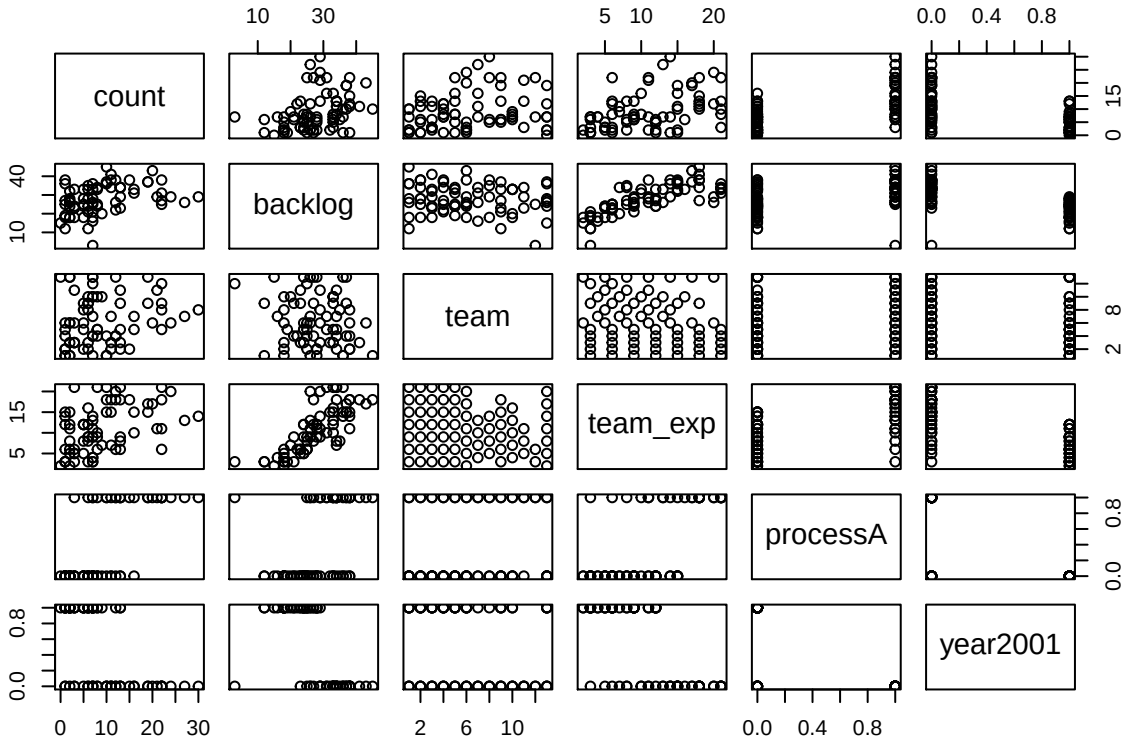
Getting to Know the Data

```
# view dataset
summary(df)
```

```
##      count           backlog           team       team_exp
##  Min.   : 0.000   Min.   : 3.00   1      : 7   Min.   : 2.00
##  1st Qu.: 3.000   1st Qu.:23.00   2      : 7   1st Qu.: 6.00
##  Median : 7.000   Median :28.00   3      : 7   Median :11.00
##  Mean   : 9.041   Mean   :27.82   4      : 7   Mean   :10.85
##  3rd Qu.:13.000   3rd Qu.:34.00   5      : 7   3rd Qu.:15.00
##  Max.   :30.000   Max.   :45.00   6      : 7   Max.   :21.00
##                                   (Other):31
##     processA         year2001
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.3562   Mean   :0.4795
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

As shown above the data has now been recoded to include the appropriate data types and is ready for futher analysis. The team categorical variable has been coerced into a factor to avoid creating a series of dumby variables.

Checking for Linearity of Variables



The scatterplot matrix plot is a useful tool for looking for linear relationship among the variables in a data set under investigation. Below is a table describing the relationships between the predictor and outcome variables.

Variable	Description
backlog	appears to be weakly positively linear with count
team	Difficult to tell due to the many different (13) team categories
team_exp	appears to be weakly positively linear with count
processA	appears to be positively linear with process A more productive
year2001	appears linear with year 2001 being less productive

Fitting the full Model

$$count = \beta_0 + \beta_1 backlog + \beta_2 team + \beta_3 team_e xp + \beta_4 process + \beta_5 year + \beta_6 quarter + \varepsilon$$

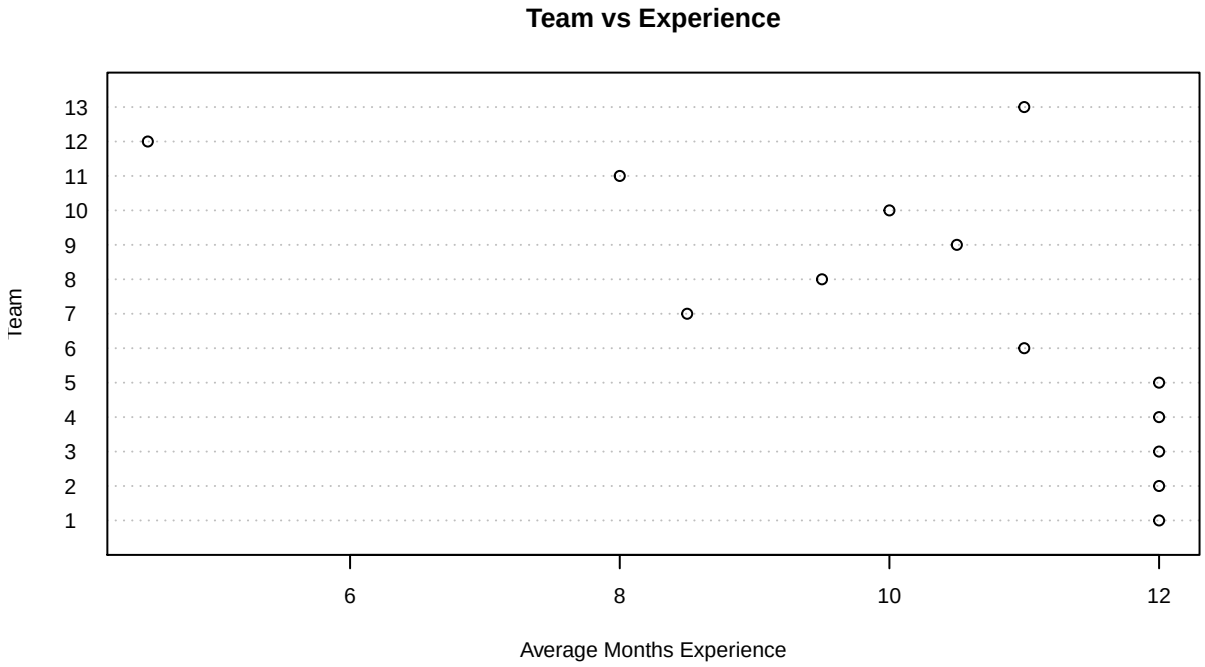
```
##
## Call:
## lm(formula = count ~ backlog + team + team_exp + processA + year2001,
##     data = df)
##
```

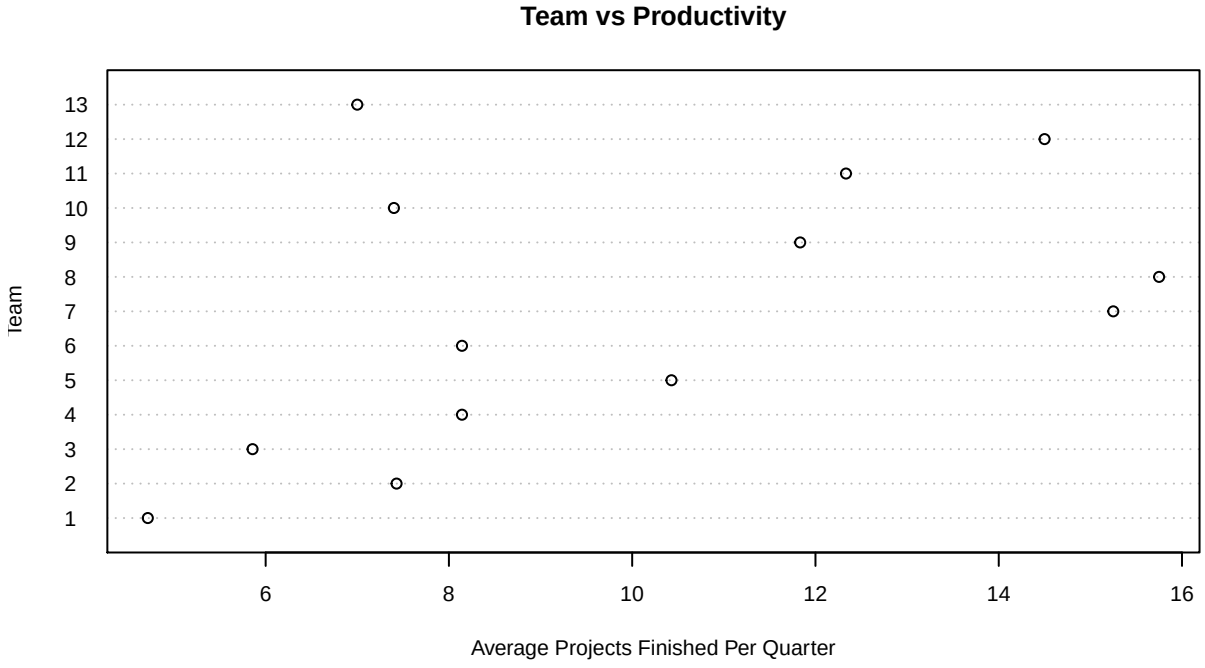
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8046 -3.0768 -0.7148  3.6123 10.0422
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.7687     5.7100  -1.010  0.31670
## backlog       0.1475     0.1400   1.053  0.29675
## team2         2.5457     2.6434   0.963  0.33966
## team3         1.1639     2.6386   0.441  0.66083
## team4         3.3654     2.6392   1.275  0.20752
## team5         5.9461     2.6477   2.246  0.02868 *
## team6         3.6087     2.6519   1.361  0.17903
## team7        10.6644     3.5105   3.038  0.00361 **
## team8        10.4628     3.3940   3.083  0.00318 **
## team9         7.7179     2.8262   2.731  0.00843 **
## team10        2.8233     3.0334   0.931  0.35598
## team11        6.8459     4.1644   1.644  0.10580
## team12        9.3288     5.6816   1.642  0.10621
## team13        2.6765     2.6505   1.010  0.31692
## team_exp      0.2222     0.2505   0.887  0.37872
## processA      7.8749     2.0775   3.791  0.00037 ***
## year2001      2.3290     2.6552   0.877  0.38417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.936 on 56 degrees of freedom
## Multiple R-squared:  0.6223, Adjusted R-squared:  0.5144
## F-statistic: 5.767 on 16 and 56 DF,  p-value: 4.092e-07
```

The call to the summary function of the lm object displays the following:

- Summary statistics of residuals
- the OLS estimate of β_j coefficients of the model
- Error estimates and associated p values to assess statistical significance of the parameters based of t-tests where:
- $H_o : \beta_j = 0 \quad where \quad p - value > 0.05$
- $H_a : \beta_j <> 0 \quad where \quad p - value \leq 0.05$
- Multiple R-squared tells the proportion of variance in the outcome variable explained by the model
- Adjusted R-squared is a more robust version of R-squared that accounts for overfitting by adding variables
- F-statistic and p-value which assesses the statistical significance of the model as a whole
- $H_o : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0 \quad where p - value > 0.05$
- $H_a : \beta_j <> for \quad atleast \quad one \quad j = 1, 2, ..., p-1 \quad where \quad p - value \quad \leq 0.05$

A rough look at the initial t-tests of the full model leads me to believe that there are differences in productivity among the teams. Teams 5, 7, 8, and 9 appear to have a statistically significant impact on productivity. However, the team experience as a whole does not appear to significant for the model. I am a bit skeptical of this because it seams like teams that have more experience working together should be more productive so, I would like to view the months of experience for each team to see if the teams with significant impact in the model all have appreciably more experience than the others. I will do so with a dotchart.





The plots show that teams 1-5 have the most experience but, they have relatively low productivity. Also, it does not appear that the average months experience for the significant teams are customering or significantly different from the others. With this information I think it is safe to remove the team experience variable from the model.

Fitting the Model without Team Experience

$$count = \beta_0 + \beta_1 backlog + \beta_2 team + \beta_3 process + \beta_4 year + \beta_5 quarter + \varepsilon$$

```
##
## Call:
## lm(formula = count ~ backlog + team + processA + year2001, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2577 -2.9311 -0.4346  3.6409 10.3334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.9400     5.3151  -0.741  0.46157
## backlog       0.1865     0.1327   1.405  0.16551
## team2         2.5012     2.6380   0.948  0.34705
## team3         1.1695     2.6337   0.444  0.65868
## team4         3.3487     2.6342   1.271  0.20881
## team5         6.0073     2.6418   2.274  0.02676 *
## team6         3.3753     2.6339   1.281  0.20521
## team7         9.4302     3.2171   2.931  0.00485 **
## team8         9.3242     3.1362   2.973  0.00431 **
## team9         7.3913     2.7969   2.643  0.01060 *
## team10        2.1458     2.9302   0.732  0.46698
## team11        5.0114     3.6082   1.389  0.17027
## team12        6.9545     5.0027   1.390  0.16989
## team13        2.4988     2.6380   0.947  0.34751
## processA      8.8750     1.7420   5.095 4.12e-06 ***
## year2001      1.3378     2.4043   0.556  0.58010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.927 on 57 degrees of freedom
## Multiple R-squared:  0.617,  Adjusted R-squared:  0.5162
## F-statistic: 6.122 on 15 and 57 DF,  p-value: 2.23e-07
```

Note that the R squared value changed very little between the two models which further lend to the argument that team experience had any effect on the model. It is still worth noting that the only other quantitative variable, backlog count, has no statistical effect on the model with a pvalue of 0.4615 which tells me it too should be removed from the model. At this point all the variables are categorical the use of linear regression for predicting productivity is probably no longer appropriate.

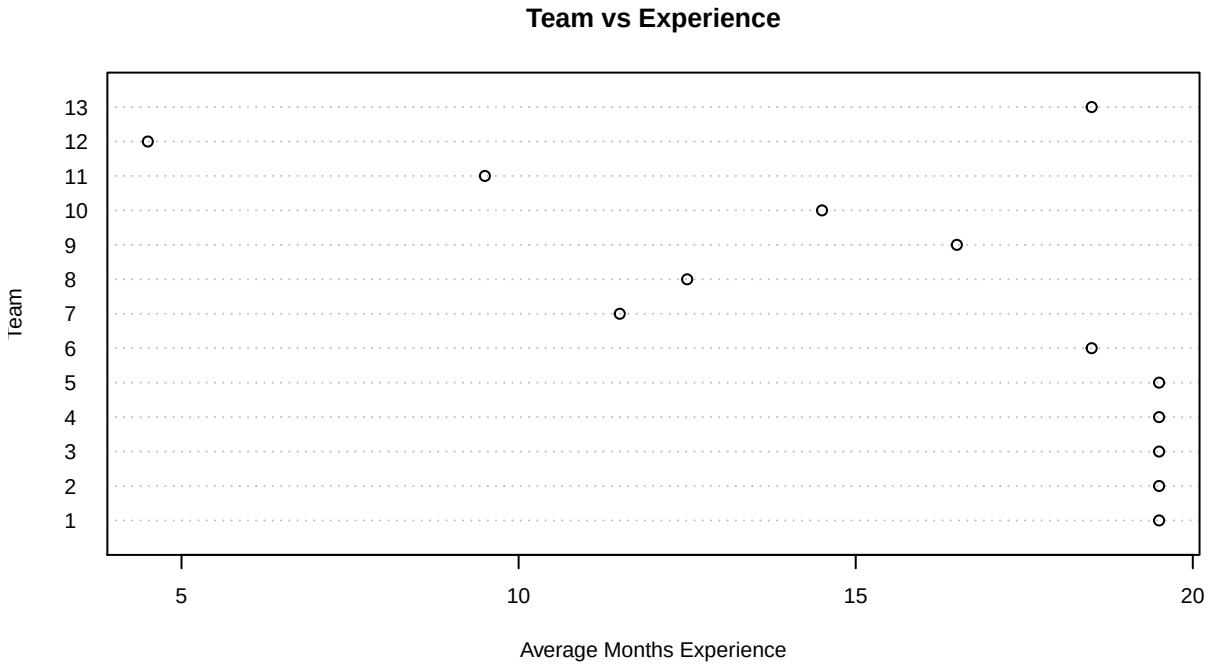
However, it should be noted that it is apparent that the new process enacted in the second year has a significant possitive effect on the productivity of the model. Another interesting aspect of the data set to examine will be whether the overall productivity of the teams smoothed out when the company switched to the new process. To do this I will analyze only the data collected on the teams using the new process.

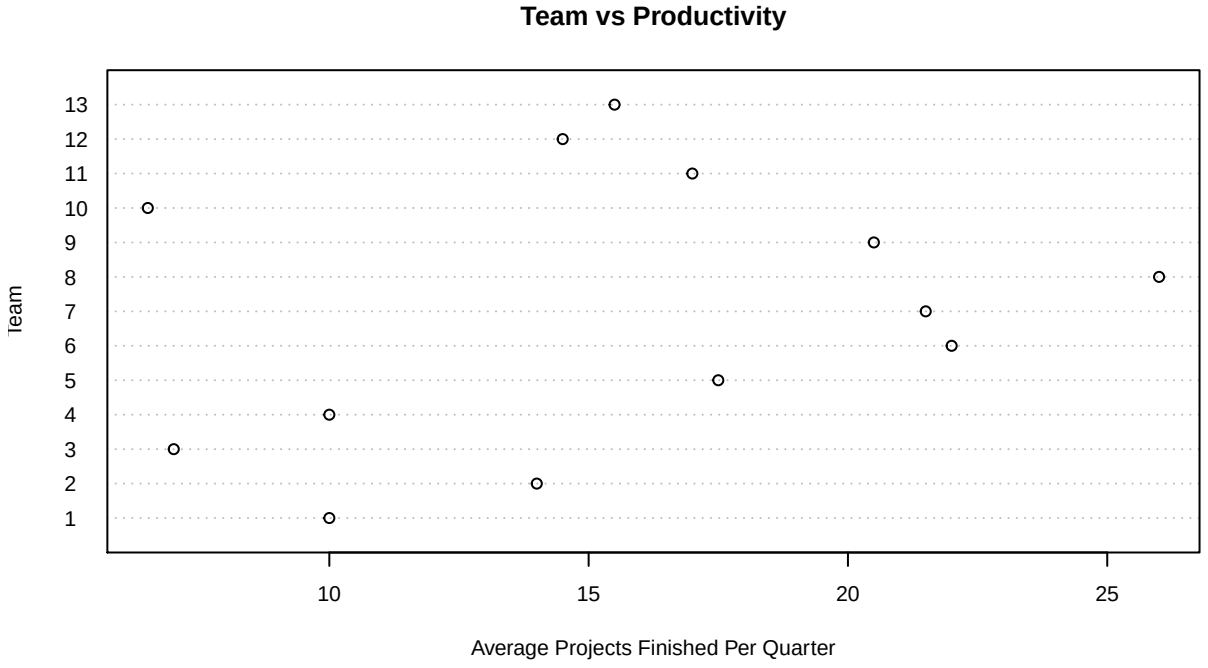
Analyzing New Process Data

```
##
## Call:
## lm(formula = count ~ ., data = processNew)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.699 -2.764  0.000  2.764  4.699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.5538    19.1617  -1.908  0.08287 .
## backlog       0.4050     0.2108   1.921  0.08095 .
## team2         5.8226     4.6697   1.247  0.23833
## team3        -1.5824     4.6314  -0.342  0.73904
## team4         1.8226     4.6697   0.390  0.70376
## team5        10.7402     4.8734   2.204  0.04975 *
## team6        15.3688     4.7837   3.213  0.00826 **
## team7        28.7297     8.3684   3.433  0.00559 **
## team8        29.6584     7.2419   4.095  0.00177 **
## team9        18.5812     5.6360   3.297  0.00712 **
## team10        7.2686     6.3321   1.148  0.27536
## team11       26.1070     9.1138   2.865  0.01539 *
## team12       38.4257    14.5092   2.648  0.02265 *
## team13       10.6914     5.1184   2.089  0.06076 .
## team_exp      1.5462     0.6801   2.273  0.04404 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.572 on 11 degrees of freedom
## Multiple R-squared:  0.8131, Adjusted R-squared:  0.5752
## F-statistic: 3.418 on 14 and 11 DF,  p-value: 0.02348
```

The pvalues for the model's β_j parameters are now showing that the team experience variable is significant at $\alpha = 0.05$ and now the only teams that are not significant are teams 13, 10, 1, 2, 3 and 4. Backlog still is not significant and should be dropped. The results indicate that for this data set the teams appear to have similar significance on the relationship on productivity but it is not an appropriate dataset for linear regression. I will again look for a recognizable relationship between the teams with significant effects on the model to months experience.





Again, there is not a clear relationship between the statistially significant teams and the average months of experience for a given team.

Result and Conclusion

The results of this regression analysis indicate that the new process for software development enacted in the second year (year 2002) had the most significant impact on quarterly project productivity. The reduced model where months of experience is removed shows that the effect of switching to the new process resulted in an increase of about 9 completed projects a quarter where all other variables are held constant.

Analysis of the data collected for only the new process resulted in a reduced dataset of 73 orginal observations to just 26 observations. Removing the extreme variation introducted by the new process showed that there was less variation among the individual teams and a greater significance on the months of experience for teams and the impact on productivity. However, due to the large number of different number of team categories and lack of significant quantitative variables using a linear regression techique is limited in the ability to draw additional significant conclusions for this data set. Perhaps a more appropriate method of analysis would be done using design of experiment or other multivariate techniques.

References

1. EMC Education Services. Data Science and Big Data Analytics. (2015). Wiley Publishing.
2. Kutner, Nachtsheim, and Neter, Applied Linear Regression Models. (2004). The McGraw-Hill Companies.4th Edition.