Search for Cause III

1 Question

The objective of this investigation is to determine which if any, of the suspects is a dominant cause. From our previous investigation, we determined the dominant cause resides in component D. This makes our suspects x10, x11 and x12 which are all continuous varying inputs.

2 Plan

To investigate our suspects x10, x11 and x12, we will perform a group comparison analysis which requires a retrospective investigation. One group will contain 8 parts which had the highest y100 values from our y100 baseline investigation and the other group will contain 8 parts with the lowest y100 values. We will measure the varying inputs x10, x11 and x12 of those parts. The total cost will be $304.

Next we will investigate the input/output relationship. Another baseline investigation will be performed measuring output y100 and varying inputs x10, x11 and x12. We sampled 35 parts over 15 shifts (5 days) using random and systematic sampling costing $8400

3 Data

3.1 Y100 Baseline Investigation Data

Daycount	Shift	Partnum	yr100
1	16	21637	-3.2
2	16	21641	-3.6
3	16	21689	0.4
4	16	21690	1.8
5	16	21697	-6.8
6	16	21698	-7.6

Table 1: View of Y100 Baseline Data

3.2 Top 8 Largest and Smallest

Rank	Partnum	yr100
1	27675	8.2
2	27208	9.6
3	25299	8.0
4	25777	7.8
5	24742	8.0
6	24226	8.4
7	23809	7.8
8	22410	8.2

Table 2: Top 8 Largest

Rank	Partnum	yr100
1	28147	-10.8
2	27897	-12.4
3	27127	-9.2
4	26035	-8.2
5	24810	-8.6
6	24252	-8.2
7	23193	-9.0
8	23106	-8.6

Table 3: Top 8 Smallest

3.3 Retrospective Dataset

Daycount	Shift	Partnum	yr100	x10	x11	x12
1	16	22410	8.2	1.8	11.8	16.8
2	17	23106	-8.6	4.6	0.5	14
3	17	23193	-9.0	1.6	7.2	-1.7
4	17	23809	7.8	3.2	-0.4	8.4
5	17	24226	8.4	1.5	8.5	13.2
6	17	24252	-8.2	3.8	3.1	3.3

Table 4: View of Retrospective Dataset

3.4 Input/Output Dataset

Daycount	Shift	Partnum	y100	x10	x11	x12
1	25	35538	-7.8	2.3	7.1	-2.6
2	25	35542	-6.2	2.8	10.6	3.4
3	25	35549	2.6	5.9	-3.6	14.4
4	25	35553	-3.8	3.6	2.2	0.0
5	25	35556	2.6	2.6	8.6	10.1
6	25	35573	3.0	1.8	4.9	8.5

Table 5: View of Input/Output Dataset

4 Analysis

4.1 Scatterplots of Suspects By Group

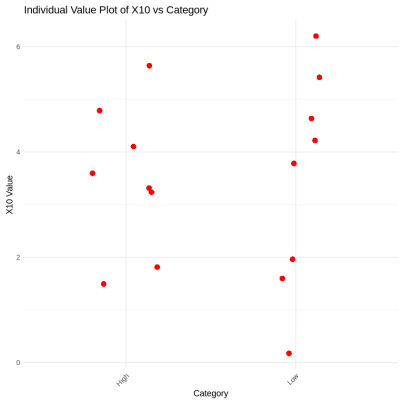


Figure 1: Individual Value Plot of X10 vs Category

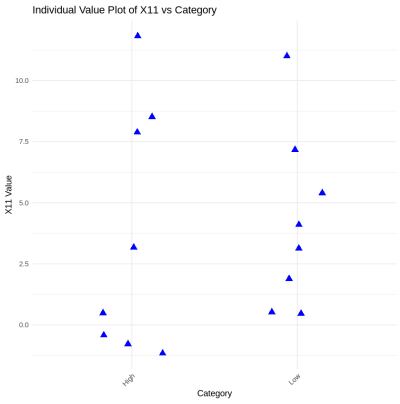


Figure 2: Individual Value Plot of X11 vs Category

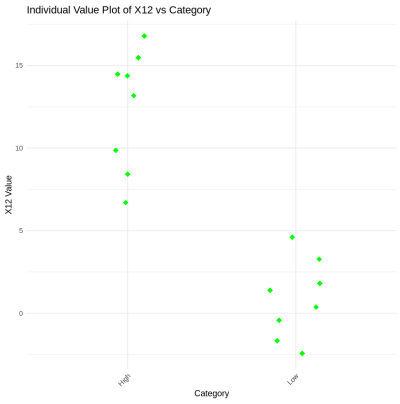


Figure 3: Individual Value Plot of X12 vs Category

The X10 plot displays a relatively tight clustering of values in both the "High" and "Low" categories. There's no apparent trend or significant difference in the spread of values between the two categories, suggesting that X10 might not be the dominant cause of variation between the "High" and "Low" categories.While the X11 plot displays more variation, it is fairly similar with no apparent pattern.

The X12 plot shows a wide spread of values across both categories. However, there is a discernible descending pattern from "High" to "Low." The values in the "High" category seem to start higher and reduce towards the "Low" category. This pattern suggests that X12 may be a significant cause to variation.

4.2 Linear Model

```
Call:
lm(formula = y100 ~ x10 + x11 + x12 + x10 * x11 * x12, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7452  -1.9722   0.0741   2.0391   9.4356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.0992447  1.1063533  -5.513 5.58e-08 ***
x10          0.4837679  0.2554820   1.894   0.0588 .
x11          0.1580866  0.1195551   1.322   0.1867
x12          0.9415486  0.1631242   5.772 1.35e-08 ***
x10:x11     -0.0152253  0.0303403  -0.502   0.6160
x10:x12     -0.0359921  0.0375911  -0.957   0.3388
x11:x12     -0.0184302  0.0167759  -1.099   0.2724
x10:x11:x12  0.0008494  0.0041820   0.203   0.8391
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.01 on 517 degrees of freedom
Multiple R-squared:  0.5029,    Adjusted R-squared:  0.4962
F-statistic: 74.72 on 7 and 517 DF,  p-value: < 2.2e-16
```

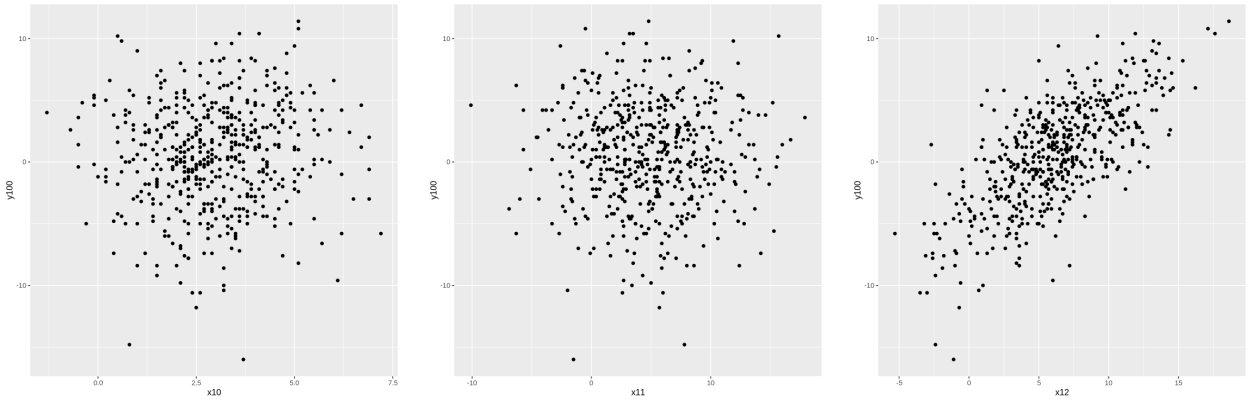
This output presents the results from a linear regression analysis using R, with ‘y100‘ as the dependent variable modeled by independent variables ‘x10‘, ‘x11‘, ‘x12‘, and their interaction terms. The R-squared value of the model is 0.5029, indicating that approximately 50.29% of the variability in ‘y100‘ can be explained by the combination of these variables and their interactions. This is a moderate amount of variance explained, suggesting the model has a reasonable fit to the data.

When examining the significance of individual variables, ‘x12‘ shows a strong and statistically significant effect on ‘y100‘, with a p-value of 1.35e-08. This suggests that ‘x12‘ is a strong predictor of the outcome variable ‘y100‘. In contrast, ‘x10‘ has a p-value of 0.0588, which is marginally above the common alpha level of 0.05, indicating a suggestive but not conclusively significant relationship with ‘y100‘. The variable ‘x11‘ has a p-value of 0.1867, making it not statistically significant.

The interaction terms in the model (‘x10:x11‘, ‘x10:x12‘, ‘x11:x12‘, and ‘x10:x11:x12‘) do not show statistical significance since all their p-values are well above the 0.05 threshold. This implies that the interactions between these variables do not significantly contribute to the model’s ability to predict ‘y100‘.

Finally, the F-statistic of 74.72 with a very small p-value suggests that the overall regression model is statistically significant and that the included variables, as a whole, have a strong association with ‘y100‘. However, given that some individual predictors and interaction terms are not significant, further model refinement or the exploration of non-linear relationships might be warranted.

4.3 Scatterplots of Each Suspect



The examination of the scatter plots for ‘x10‘ and ‘x11‘ against the dependent variable ‘y100‘ does not reveal any discernible patterns, indicating that these variables might not have a strong linear relationship with ‘y100‘. Conversely, the scatter plot for ‘x12‘ demonstrates a pronounced linear trend with ‘y100‘. This observation strongly suggests that ‘x12‘ plays a significant role in explaining the variation in ‘y100‘.

5 Conclusion

The residual standard error (RSE) from the linear model is 3.01, which is substantially lower than the baseline standard deviation of 4.28 for the ‘y100‘ values. The RSE measures the typical size of the residuals and, hence, how well the model’s predictions match the observed data. When the RSE is significantly smaller than the standard deviation of the observed values, it indicates that the model has added explanatory power relative to the simple mean of the data. In this context, the reduction in variability from 4.28 to 3.01 suggests that the model captures a considerable amount of the variance in the outcome variable, which could not be explained merely by chance.

Given this reduction in variance, it can be inferred that at least one of the independent variables in the model is providing substantial explanatory power. The analysis has identified ‘x12‘ as this variable, based on its statistical significance in the model and the clear linear trend observed in the scatter plots. Consequently, ‘x12‘ can be considered a dominant factor influencing the variations in ‘y100‘.’