Accident Analysis for Assessment of Positive Train Control

Omkar G. Kharkar

Honor Pledge: On my honor, I pledge that I am the sole author of this paper and I have accurately cited all help and references used in its completion

Summary:

In this report, I have included an analysis of the Positive Train Control technology and an analysis behind why the federal government should continue funding this technology only if the sole focus to reduce the number of injuries and reduce the cost of damage. My analysis has been based on statistical analysis and linear models of the data that best explain trends in rail accidents seen in the year 2010.

1 Problem Description

1.1 Background

The purpose of this report to evaluate the effectiveness of Positive Train Control (hereby abbreviated frequently as PTC) by assessing whether this technology is capably of reducing the frequency of train accidents as well as their severity. From [3], Brown, D. E and Wang, X. (2011) provide a more detailed description about this study:

"The federal government wants to reduce the number and severity of train accidents. Positive Train Control (PTC) is a suite of technologies proposed to achieve this goal. Earlier in the decade, PTC was among the top 10 most wanted" safety improvements by the National Transportation Safety Board (NTSB) to improve transportation safety and at least one of the Federal Rail-road Administration (FRA) web sites still calls it that [1]. However, the current NTSB top 10 most wanted improvements list does not include PTC [2]. How important is PTC for train safety?

This report provides an assessment of the importance of PTC in reducing the number and severity of train accidents. This assessment can inform FRA planning for the future of PTC. The need to conduct this particular study is given in [3].

PTC involves the integration and use of multiple technologies (e.g., accelerometers, controllers, temperature, humidity, and other environmental sensors, and GPS). These technologies must operate seamlessly with people (e.g., train engineers and dispatchers) and organizations (e.g., railroads, FRA, and unions). The FRA describes PTC as follows:

> 'Positive Train Control (PTC) refers to technology that is capable of preventing train-to-train collisions, overspeed derailments, and casualties or injuries to roadway workers (e.g., maintenance-of-way workers, bridge workers, signal maintainers) operating within their limits of authority as a result of unauthorized incursion by a train. PTC is also capable of preventing train movements through a switch left in the wrong position. PTC systems vary widely in complexity and sophistication based on the level of automation and functionality they implement, the system architecture utilized, the wayside system upon which they are based (i.e., non-signaled, block signal, cab signal, etc.), and the degree of train control they are capable of assuming. [4]'

… While the FRA has contractual mechanisms to spend $50M per year through 2013 to continue PTC deployments the funds for this have not yet been allocated. Before they are we need to fully understand the potential safety contributions from PTC." (p.1)

- D.E Brown & X. Wang, "Laboratory 1: Positive Train Control", August 2011, assignment in class SYS 4021

1.2 Goals

The overall goal of this study was to understand if PTC can reduce the number and frequency of train accidents, which defined the metrics of interest here.

Goal:
To determine whether Positive Train Control can significantly reduce the number and severity of train accidents

- To determine whether Positive Train Control can significantly minimize the severity from train accidents
    - To determine whether Positive Train Control can significantly reduce property damage (measured in U.S dollars $)
    - To determine whether Positive Train Control can significantly reduce casualties and loss of human life (measured by a combination of total injuries and casualties)
- To determine whether Positive Train Control can significantly minimize the frequency of train accidents

Note: The phrase "significantly" here is used in the statistical context, under which the hypotheses (stated below) were tested

1.3 Data Source

The major source of accident data that was used for this study comes from [5]. The figures and statistics computed were done so by using Rail Accident data that was obtained from the FRA for the year 2010. It should be noted that further extensions to this study, including the statistical testing and modeling that were performed could be extended to prior years, provided that this data would be available. This would allow for a more holistic approach and could lead to further insight on whether Positive Train Control can truly achieve the goals outlined above.

Also, it should be noted that from above, experimental evaluation of PTC vs. Non-PTC is difficult. The FRA has not done a large-scale deployment of Positive Train Control throughout the country, and as such, there is very little data for accidents that occurred under Positive Train Control. Therefore, to make up for the lack of data, the accident cause categories were divided as follows: The data was binned into two hypothetical categories. First, the accidents that occurred by human error were categorized as "PTC accidents," since theoretically, these accidents could have been prevented by use of PTC. However, the other categories were all grouped into the "Non-PTC accidents" category, since these accidents could not have been prevented even if Positive Train Control were fully implemented, at its current state. From this data grouping, the statistical analysis was carried out.

2 EISE Approach

2.1 Hypotheses

For this study, I conjectured three separate hypotheses, based on the goals that are outlined above, that were statistically assessed using linear models and methods.

Hypothesis 1 – Severity (Damage) of Accidents

Null Hypothesis – The cost of damage for accidents involving trains under Positive Train Control trains is significantly (p-value < 0.05) lower than the cost of damage for accidents involving trains under Positive Train Control.

Alternative Hypothesis - The cost of damage for accidents involving trains under Positive Train Control trains is not significantly (p-value < 0.05) lower than the cost of damage for accidents involving trains not under Positive Train Control.

Hypothesis 2 – Severity (Human Life) of Accidents

Null Hypothesis – The total injuries and casualties for accidents involving trains under Positive Train Control trains is significantly (p-value < 0.05) lower than the total injuries and casualties for accidents involving trains under Positive Train Control.

Alternative Hypothesis - The total injuries and casualties for accidents involving trains under Positive Train Control trains is not significantly (p-value < 0.05) lower than the total injuries and casualties for accidents involving trains not under Positive Train Control.

Hypothesis 3 – Frequency of Accidents

Null Hypothesis – The number of accidents involving trains under Positive Train Control trains is significantly (p-value < 0.05) lower than the number of accidents involving trains under Positive Train Control.

Alternative Hypothesis - The number of accidents involving trains under Positive Train Control trains is not significantly (p-value < 0.05) lower than the number of accidents involving trains not under Positive Train Control.

2.2 Visualization and Graphical Analysis

For this study, I used several univariate and multivariate graphical methods for analysis . In order to gain an understanding of the range of values and understand the data, the box plot was used, to understand whether there was a high number of accidents present, and how they were categorized. The second univariate type of graph that was used was the QQ-plot. This plot was used in conjunction with the linear models to best assess the fit of the model, and verify the assumption that the residual values are normally distributed with a distribution of N(0,1). There were also several multivariate displays that were used for this report. First, the most important display was the Scatter Plot Matrix. This matrix displays the correlations between variables, and was extremely important for this study. By observing correlations between variables, this can indicate what overlap they have, and accordingly, I could choose which variables should be included in any regression model, and which variables may overlap with another and show similar statistical significance for testing purposes. Second, principal components analysis was also used to understand the variance between the variables, and more importantly, understand the relative relationship between variables. Principal Components was very useful in identifying the response variables that were used on constructing the linear models that were related to the respective hypothesis that was being tested.

2.3 Linear Statistical Models

There were two major types of linear models that were used in this study. First, a general multi-variable linear regression was used in order to construct a hypothetical model between a given response variable(s) and a corresponding set of predictors. Second, stepwise regression was used in order to best answer the question of which variables to include in devising a model that has the best fit. To understand which models were appropriate in understand correlation between factors, I used the t-test for coefficients; to compare two different models, I used the Partial F-test and ANCOVA, along with ANOVA in some cases. When categorical variables used (ex. Weather, Cause, Visibility), these variables were coded into specific categories, that functioned as “levels.” For example, in the case of Cause (PTC vs. Non-PTC, the variables were coded as 1 or 0 respectively.) Finally, to assess the models, several measures were used, such as adjusted R-squared, AIC, BIC and diagnostic plots. The models did not have any severe departures from the assumptions that were made and as a result, they did not require transformations, such as box-cox or second-order models.

3 Evidence

3.1 Visualization and Graphical Analysis Results

Based on the application of Box plots applied to several aspects, such as injuries, and cost of damage, I noted that the number of accidents that occur as a result of Non-PTC related data is far greater than the number of accidents that occur when PTC is actually present. As a result, I rejected the hypothesis that PTC can significantly reduce the number of accidents that are caused.

The Scatter Plot Matrices (enclosed) show the correlation between variables and they demonstrate the relationships that are present in such a case. There are no alarmingly high correlations between variables that were chosen. When principal components analysis was used, the TOTINJ and TOTKLD attributes were identified as the logical variables of choice as response variables for assessing injury, and ACCDMG was the logical choice for the response variable for measuring cost of damage.

3.2 Linear Statistical Modeling Results

For this project, several linear models were estimated, and assessed to verify which one had the best fit for the data, and correctly factored into account the cause variable, along with other factors that best explain the relationship between the given predictors and the response variable.

This allowed me to assess what role the variable Cause played in each of these analyses. For the Cost of damage hypothesis, the null hypothesis was not rejected based on the Step-wise model 2 that was used and the results of the t-test for the cause variable. (Refer to Figure 3). However, it should be noted that the cause variable in the model was significant at the 0.1 level. Thus, further studies may need to be done to understand how sensitive the model is. For the Injury hypothesis, the null hypothesis was rejected at a 0.05 significance level, based on the t-test result for the cause variable in this model.

4 Recommendations

Based on the modeling done here, my recommendation for the FRA is to implement Positive Train Control only as a means of lowering the injuries & casualties associated with rail accidents, or for lowering the damage of accidents. It is unclear and not demonstrated here that Positive Train Control can significantly reduce the frequency of accidents.

5 References

[1] Positive train control (ptc)," Federal Railroad Administration, 2011. [Online]. Available: http://www.fra.dot.gov/us/content/784.shtml

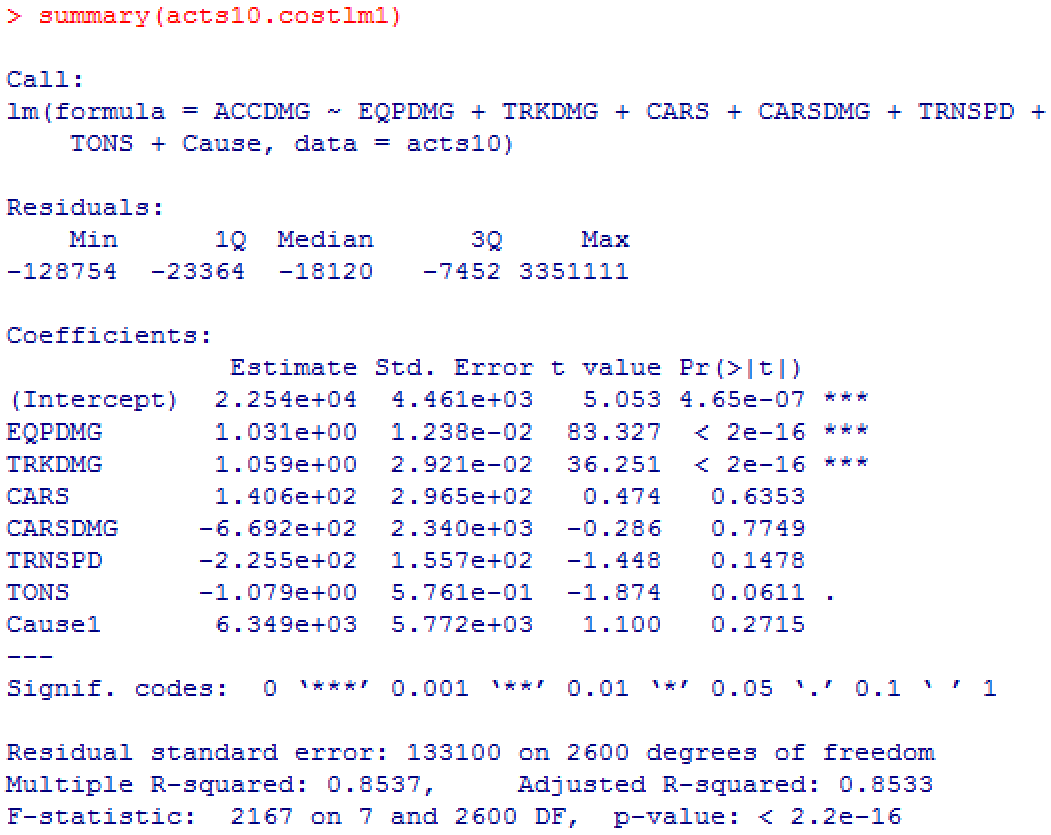
[2] Most wanted list," National Transportation Safety Board. [Online]. Available: http://www.ntsb.gov/safety/mwl.html

[3] D. E. Brown and X. Wang, \Laboratory 1: Positive train control," August 2011, assignment in class SYS 4021.
[4] Positive train control overview," Federal Railroad Administration, February 2009. [Online]. Available: http://www.fra.dot.gov/rrs/pages/fp 1265.shtml

[5] Home," Federal Railroad Administration O_ce of Safety Analysis, 2011. [Online]. Available: http://safetydata.fra.dot.gov/o_ceofsafety/

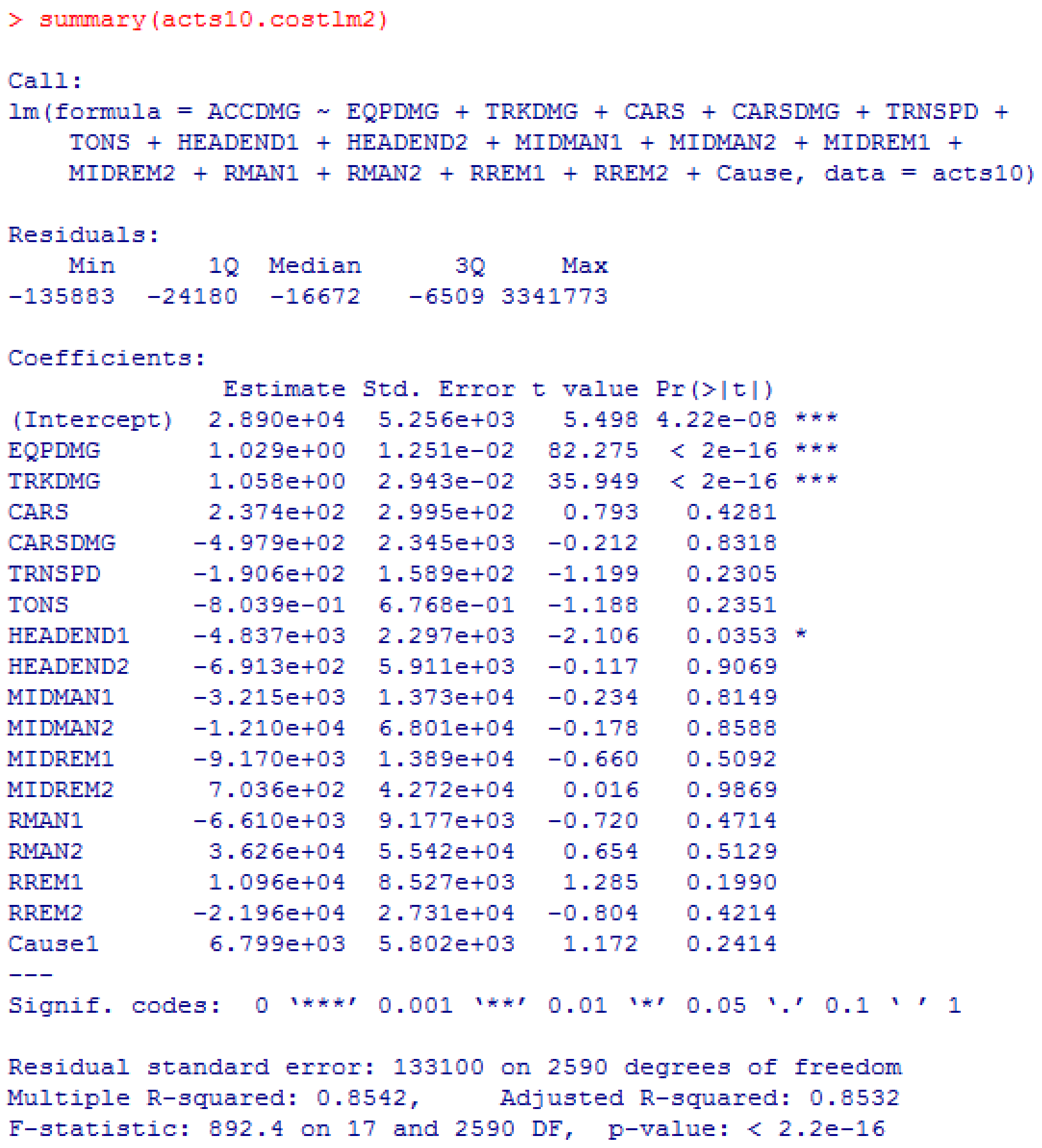
Appendix A

Cost of Damage Data



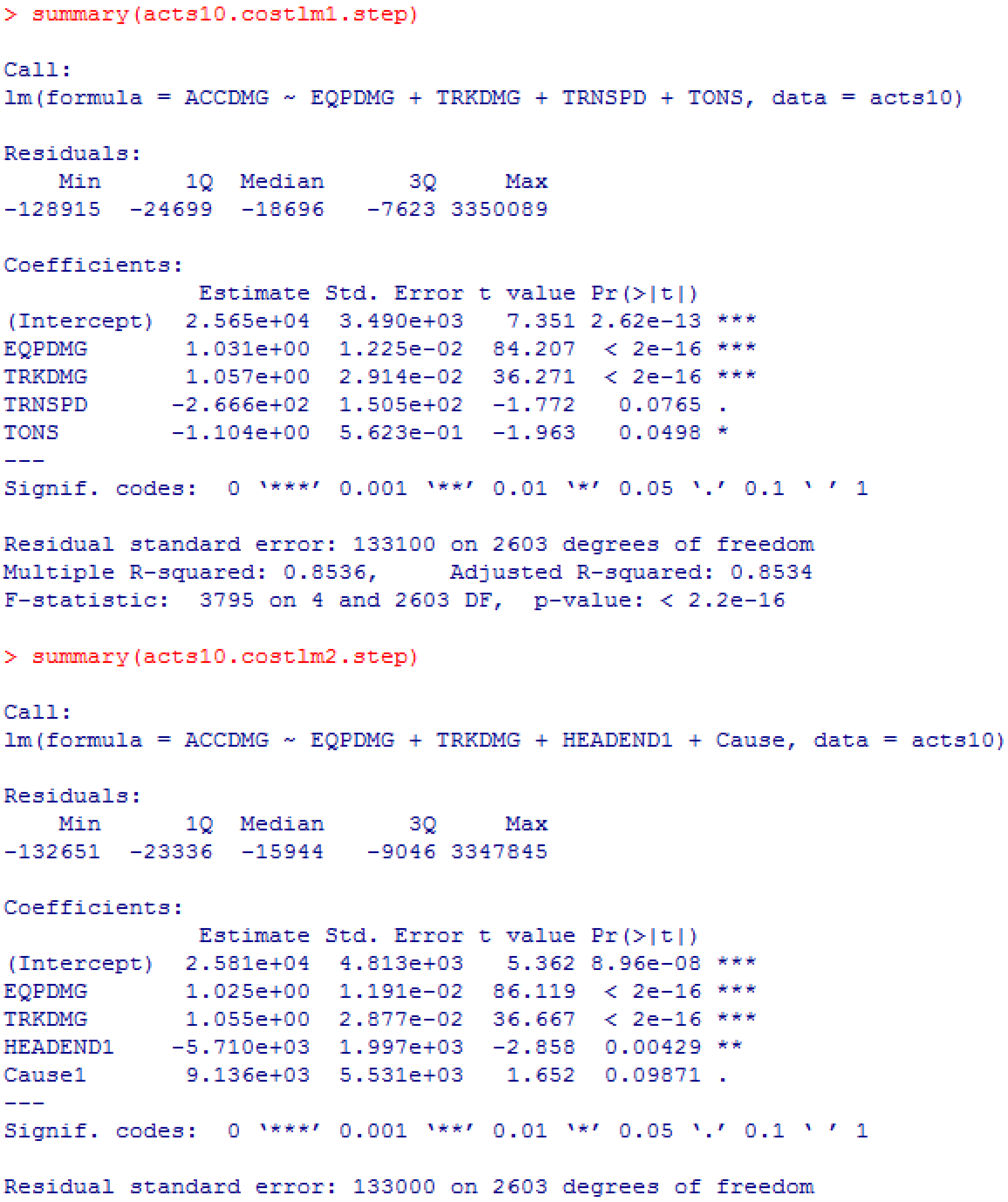
```
> summary(acts10.costlm1)

Call:
lm(formula = ACCDMG ~ EQPDMG + TRKDMG + CARS + CARSDMG + TRNSPD +
    TONS + Cause, data = acts10)

Residuals:
    Min      1Q  Median      3Q     Max
-128754  -23364  -18120   -7452 3351111

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.254e+04  4.461e+03   5.053 4.65e-07 ***
EQPDMG       1.031e+00  1.238e-02  83.327  < 2e-16 ***
TRKDMG       1.059e+00  2.921e-02  36.251  < 2e-16 ***
CARS         1.406e+02  2.965e+02   0.474   0.6353
CARSDMG     -6.692e+02  2.340e+03  -0.286   0.7749
TRNSPD      -2.255e+02  1.557e+02  -1.448   0.1478
TONS        -1.079e+00  5.761e-01  -1.874   0.0611 .
Cause1       6.349e+03  5.772e+03   1.100   0.2715
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133100 on 2600 degrees of freedom
Multiple R-squared: 0.8537,     Adjusted R-squared: 0.8533
F-statistic:  2167 on 7 and 2600 DF,  p-value: < 2.2e-16
```

Figure 1 – General Linear Model



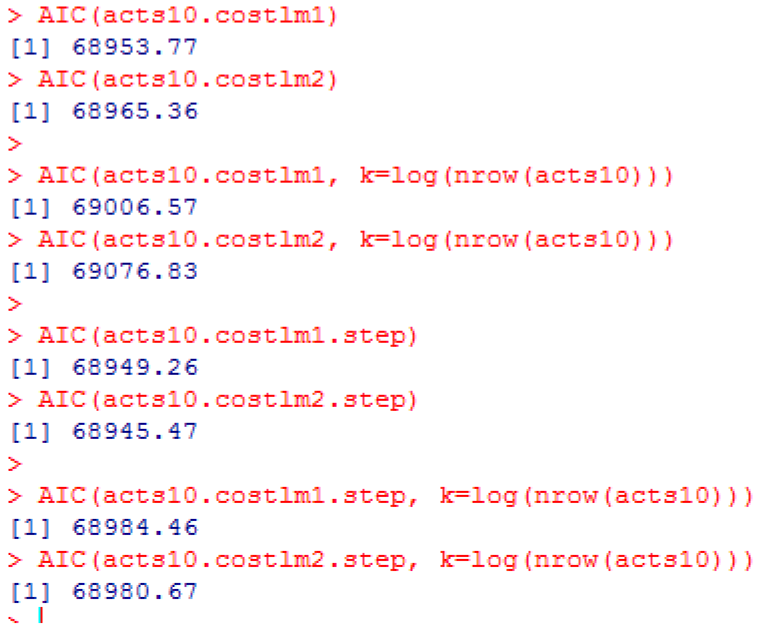
```
> summary(acts10.costlm2)

Call:
lm(formula = ACCDMG ~ EQPDMG + TRKDMG + CARS + CARSDMG + TRNSPD +
    TONS + HEADEND1 + HEADEND2 + MIDMAN1 + MIDMAN2 + MIDREM1 +
    MIDREM2 + RMAN1 + RMAN2 + RREM1 + RREM2 + Cause, data = acts10)

Residuals:
    Min      1Q  Median      3Q     Max
-135883  -24180  -16672   -6509 3341773

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.890e+04  5.256e+03   5.498 4.22e-08 ***
EQPDMG       1.029e+00  1.251e-02  82.275  < 2e-16 ***
TRKDMG       1.058e+00  2.943e-02  35.949  < 2e-16 ***
CARS         2.374e+02  2.995e+02   0.793   0.4281
CARSDMG     -4.979e+02  2.345e+03  -0.212   0.8318
TRNSPD      -1.906e+02  1.589e+02  -1.199   0.2305
TONS        -8.039e-01  6.768e-01  -1.188   0.2351
HEADEND1    -4.837e+03  2.297e+03  -2.106   0.0353 *
HEADEND2    -6.913e+02  5.911e+03  -0.117   0.9069
MIDMAN1     -3.215e+03  1.373e+04  -0.234   0.8149
MIDMAN2     -1.210e+04  6.801e+04  -0.178   0.8588
MIDREM1     -9.170e+03  1.389e+04  -0.660   0.5092
MIDREM2      7.036e+02  4.272e+04   0.016   0.9869
RMAN1       -6.610e+03  9.177e+03  -0.720   0.4714
RMAN2        3.626e+04  5.542e+04   0.654   0.5129
RREM1        1.096e+04  8.527e+03   1.285   0.1990
RREM2       -2.196e+04  2.731e+04  -0.804   0.4214
Cause1       6.799e+03  5.802e+03   1.172   0.2414
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133100 on 2590 degrees of freedom
Multiple R-squared: 0.8542,     Adjusted R-squared: 0.8532
F-statistic: 892.4 on 17 and 2590 DF,  p-value: < 2.2e-16
```

Figure 2 – Second General Linear Model – Locomotive Data



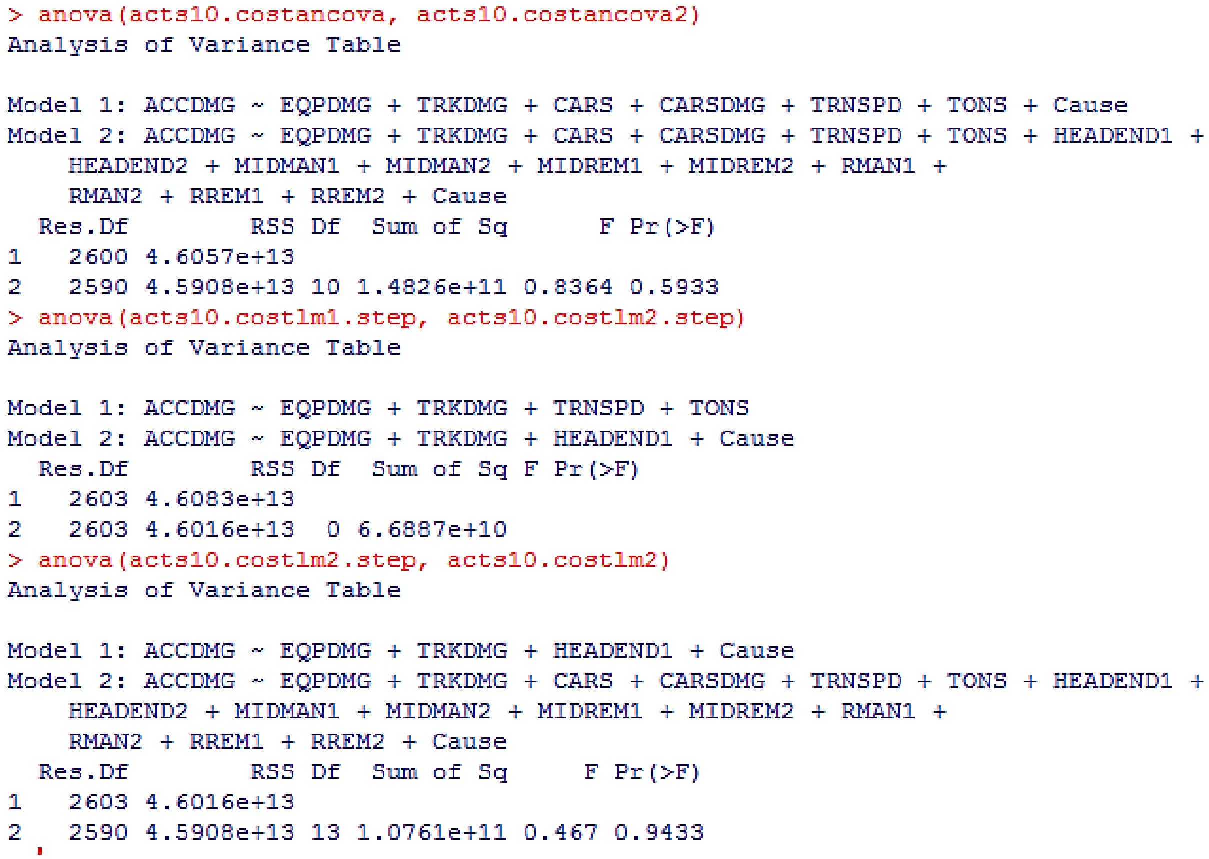
```
> summary(acts10.costlm1.step)

Call:
lm(formula = ACCDMG ~ EQPDMG + TRKDMG + TRNSPD + TONS, data = acts10)

Residuals:
    Min      1Q  Median      3Q     Max 
-128915  -24699  -18696   -7623 3350089 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.565e+04  3.490e+03   7.351 2.62e-13 ***
EQPDMG       1.031e+00  1.225e-02  84.207  < 2e-16 ***
TRKDMG       1.057e+00  2.914e-02  36.271  < 2e-16 ***
TRNSPD      -2.666e+02  1.505e+02  -1.772   0.0765 .  
TONS        -1.104e+00  5.623e-01  -1.963   0.0498 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133100 on 2603 degrees of freedom
Multiple R-squared: 0.8536,     Adjusted R-squared: 0.8534 
F-statistic:  3795 on 4 and 2603 DF,  p-value: < 2.2e-16 

> summary(acts10.costlm2.step)

Call:
lm(formula = ACCDMG ~ EQPDMG + TRKDMG + HEADEND1 + Cause, data = acts10)

Residuals:
    Min      1Q  Median      3Q     Max 
-132651  -23336  -15944   -9046 3347845 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.581e+04  4.813e+03   5.362 8.96e-08 ***
EQPDMG       1.025e+00  1.191e-02  86.119  < 2e-16 ***
TRKDMG       1.055e+00  2.877e-02  36.667  < 2e-16 ***
HEADEND1    -5.710e+03  1.997e+03  -2.858  0.00429 ** 
Cause1       9.136e+03  5.531e+03   1.652  0.09871 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133000 on 2603 degrees of freedom
```

Figure 3 – Step-Wise Regression Results for General Models



```
> AIC(acts10.costlm1)
[1] 68953.77
> AIC(acts10.costlm2)
[1] 68965.36
>
> AIC(acts10.costlm1, k=log(nrow(acts10)))
[1] 69006.57
> AIC(acts10.costlm2, k=log(nrow(acts10)))
[1] 69076.83
>
> AIC(acts10.costlm1.step)
[1] 68949.26
> AIC(acts10.costlm2.step)
[1] 68945.47
>
> AIC(acts10.costlm1.step, k=log(nrow(acts10)))
[1] 68984.46
> AIC(acts10.costlm2.step, k=log(nrow(acts10)))
[1] 68980.67
```

Figure 4 – AIC, BIC Criterion for Cost of Damage Models



```
> anova(acts10.costancova, acts10.costancova2)
Analysis of Variance Table

Model 1: ACCDMG ~ EQPDMG + TRKDMG + CARS + CARSDMG + TRNSPD + TONS + Cause
Model 2: ACCDMG ~ EQPDMG + TRKDMG + CARS + CARSDMG + TRNSPD + TONS + HEADEND1 +
    HEADEND2 + MIDMAN1 + MIDMAN2 + MIDREM1 + MIDREM2 + RMAN1 +
    RMAN2 + RREM1 + RREM2 + Cause
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1   2600 4.6057e+13
2   2590 4.5908e+13 10 1.4826e+11 0.8364 0.5933
> anova(acts10.costlm1.step, acts10.costlm2.step)
Analysis of Variance Table

Model 1: ACCDMG ~ EQPDMG + TRKDMG + TRNSPD + TONS
Model 2: ACCDMG ~ EQPDMG + TRKDMG + HEADEND1 + Cause
  Res.Df        RSS Df  Sum of Sq F Pr(>F)
1   2603 4.6083e+13
2   2603 4.6016e+13  0 6.6887e+10
> anova(acts10.costlm2.step, acts10.costlm2)
Analysis of Variance Table

Model 1: ACCDMG ~ EQPDMG + TRKDMG + HEADEND1 + Cause
Model 2: ACCDMG ~ EQPDMG + TRKDMG + CARS + CARSDMG + TRNSPD + TONS + HEADEND1 +
    HEADEND2 + MIDMAN1 + MIDMAN2 + MIDREM1 + MIDREM2 + RMAN1 +
    RMAN2 + RREM1 + RREM2 + Cause
  Res.Df        RSS Df  Sum of Sq     F Pr(>F)
1   2603 4.6016e+13
2   2590 4.5908e+13 13 1.0761e+11 0.467 0.9433
```

Figure 5 – ANOVA Results for Damage Models

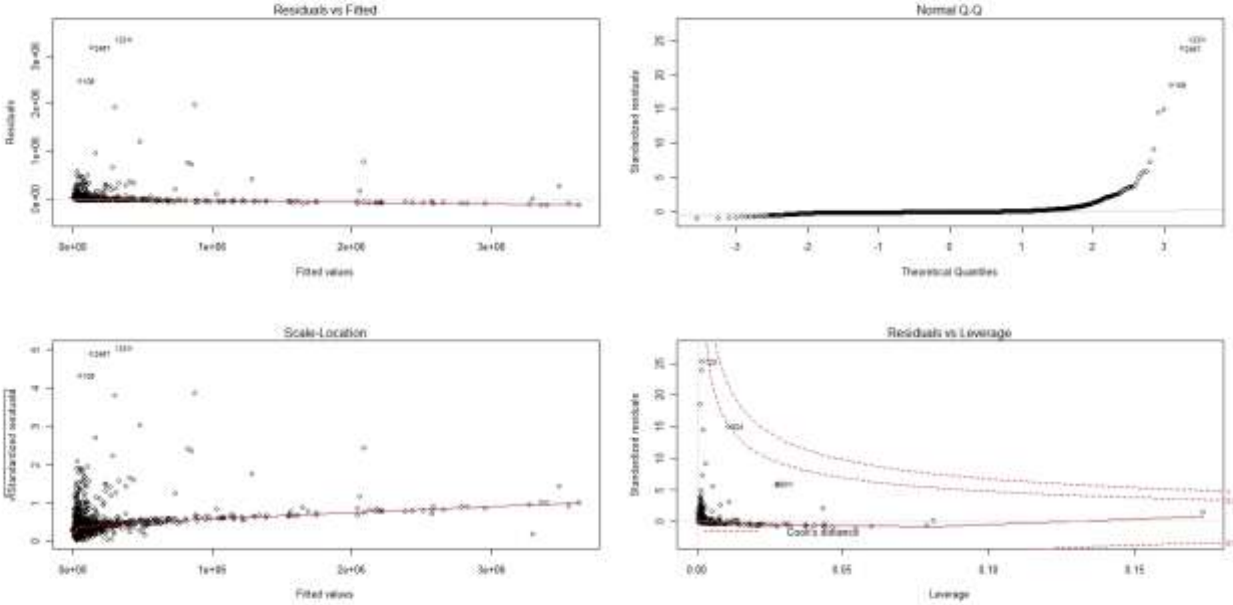
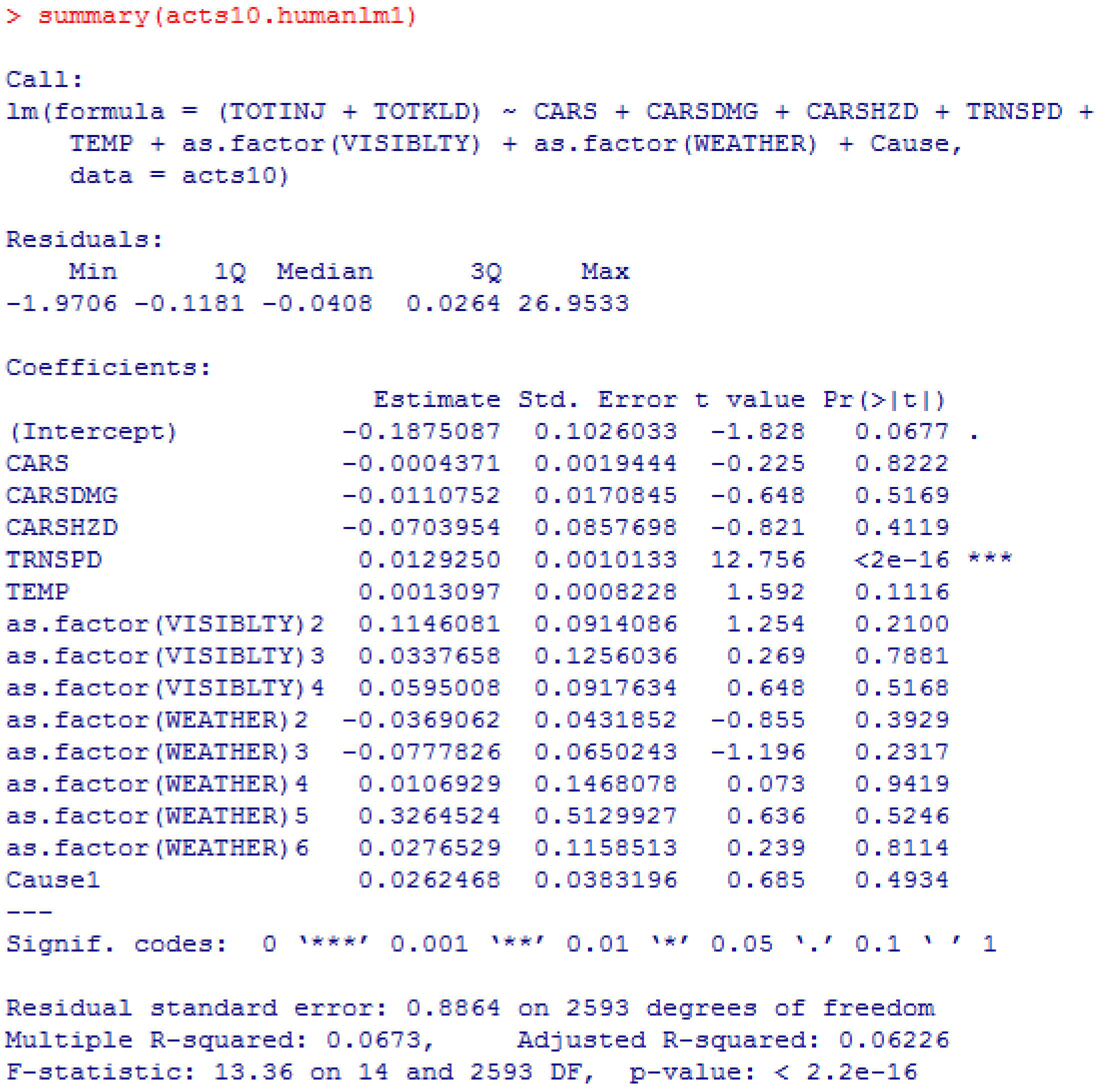


Figure 6 – Diagnostic Plot for Step-Wise Model 2 (Chosen) for cost of Damage Data

Injury Data



```
> summary(acts10.humanlm1)

Call:
lm(formula = (TOTINJ + TOTKLD) ~ CARS + CARSDMG + CARSHZD + TRNSPD +
    TEMP + as.factor(VISIBLTY) + as.factor(WEATHER) + Cause,
    data = acts10)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9706 -0.1181 -0.0408  0.0264 26.9533

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.1875087  0.1026033  -1.828   0.0677 .
CARS                 -0.0004371  0.0019444  -0.225   0.8222
CARSDMG              -0.0110752  0.0170845  -0.648   0.5169
CARSHZD              -0.0703954  0.0857698  -0.821   0.4119
TRNSPD                0.0129250  0.0010133  12.756   <2e-16 ***
TEMP                  0.0013097  0.0008228   1.592   0.1116
as.factor(VISIBLTY)2  0.1146081  0.0914086   1.254   0.2100
as.factor(VISIBLTY)3  0.0337658  0.1256036   0.269   0.7881
as.factor(VISIBLTY)4  0.0595008  0.0917634   0.648   0.5168
as.factor(WEATHER)2  -0.0369062  0.0431852  -0.855   0.3929
as.factor(WEATHER)3  -0.0777826  0.0650243  -1.196   0.2317
as.factor(WEATHER)4   0.0106929  0.1468078   0.073   0.9419
as.factor(WEATHER)5   0.3264524  0.5129927   0.636   0.5246
as.factor(WEATHER)6   0.0276529  0.1158513   0.239   0.8114
Cause1                0.0262468  0.0383196   0.685   0.4934
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8864 on 2593 degrees of freedom
Multiple R-squared: 0.0673,     Adjusted R-squared: 0.06226
F-statistic: 13.36 on 14 and 2593 DF,  p-value: < 2.2e-16
```

Figure 7 – General Model for Injury Data

```
> summary(acts10.humanlm2)

Call:
lm(formula = (TOTINJ + TOTKLD) ~ TEMP + as.factor(VISIBLTY) +
    as.factor(WEATHER) + Cause, data = acts10)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4564 -0.1813 -0.1191 -0.0445 27.7488

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.0329989  0.1038095   0.318    0.751
TEMP                  0.0010040  0.0008467   1.186    0.236
as.factor(VISIBLTY)2  0.1227956  0.0940707   1.305    0.192
as.factor(VISIBLTY)3  0.0192108  0.1292203   0.149    0.882
as.factor(VISIBLTY)4  0.0479604  0.0945044   0.507    0.612
as.factor(WEATHER)2  -0.0500248  0.0444662  -1.125    0.261
as.factor(WEATHER)3  -0.0769124  0.0669585  -1.149    0.251
as.factor(WEATHER)4   0.0526590  0.1511041   0.348    0.727
as.factor(WEATHER)5   0.2674466  0.5285467   0.506    0.613
as.factor(WEATHER)6   0.0121005  0.1189891   0.102    0.919
Cause1               -0.1067129  0.0379795  -2.810    0.005 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9134 on 2597 degrees of freedom
Multiple R-squared: 0.008115,   Adjusted R-squared: 0.004296
F-statistic: 2.125 on 10 and 2597 DF,  p-value: 0.01978
```

Figure 8 – General Model 2 for Injury Data

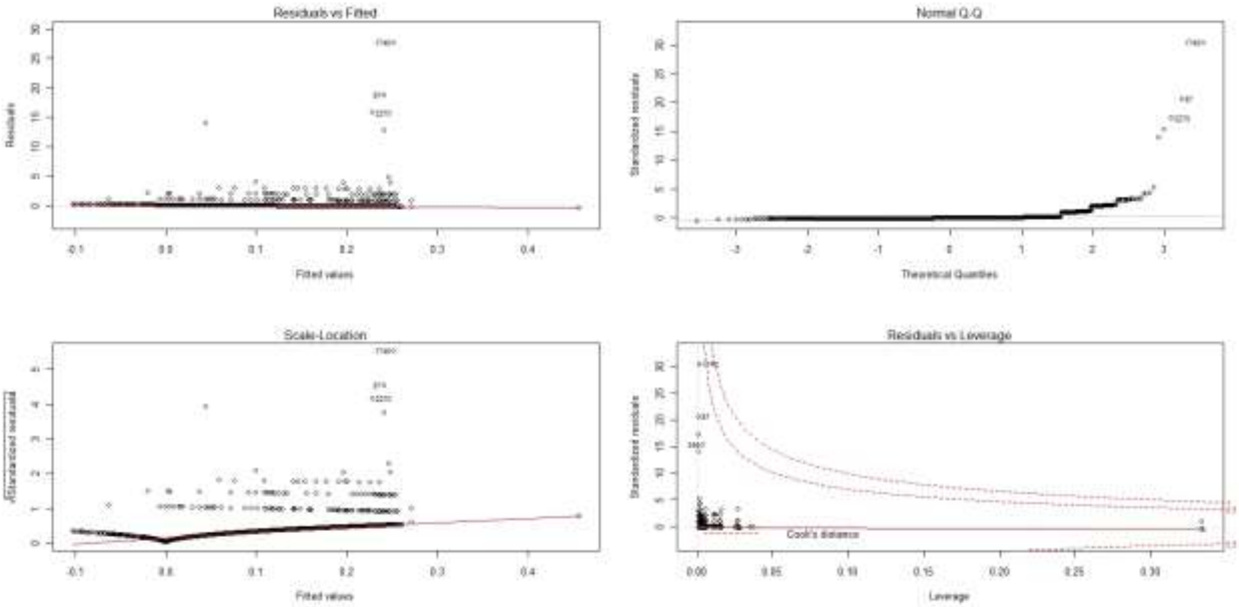


Figure 9 – Diagnostic Plot for General Linear Model 2 (Chosen)