**Regression Model**

A multiple linear regression was performed, where the fatalities (FATALS) was the dependent variable, and

STATE

COUNTY

CITY

DAY\_WEEK

NHS

FUNC\_SYS

were the independent variables.

A summary of the model was taken:

Residuals:

Min 1Q Median 3Q Max

-0.2141 -0.1028 -0.0872 -0.0722 8.8948

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.078e+00 6.230e-03 173.085 < 2e-16 \*\*\*

accident\_hour1$STATE 1.975e-04 1.332e-04 1.483 0.138190

accident\_hour1$COUNTY 8.187e-05 2.251e-05 3.637 0.000277 \*\*\*

accident\_hour1$CITY -7.056e-06 1.086e-06 -6.498 8.27e-11 \*\*\*

accident\_hour1$DAY\_WEEK 2.147e-04 9.784e-04 0.219 0.826347

accident\_hour1$NHS 6.150e-03 1.460e-03 4.213 2.53e-05 \*\*\*

accident\_hour1$FUNC\_SYS -6.243e-04 1.011e-04 -6.176 6.66e-10 \*\*\*

accident\_hour1$DRUNK\_DR 3.635e-02 4.393e-03 8.275 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

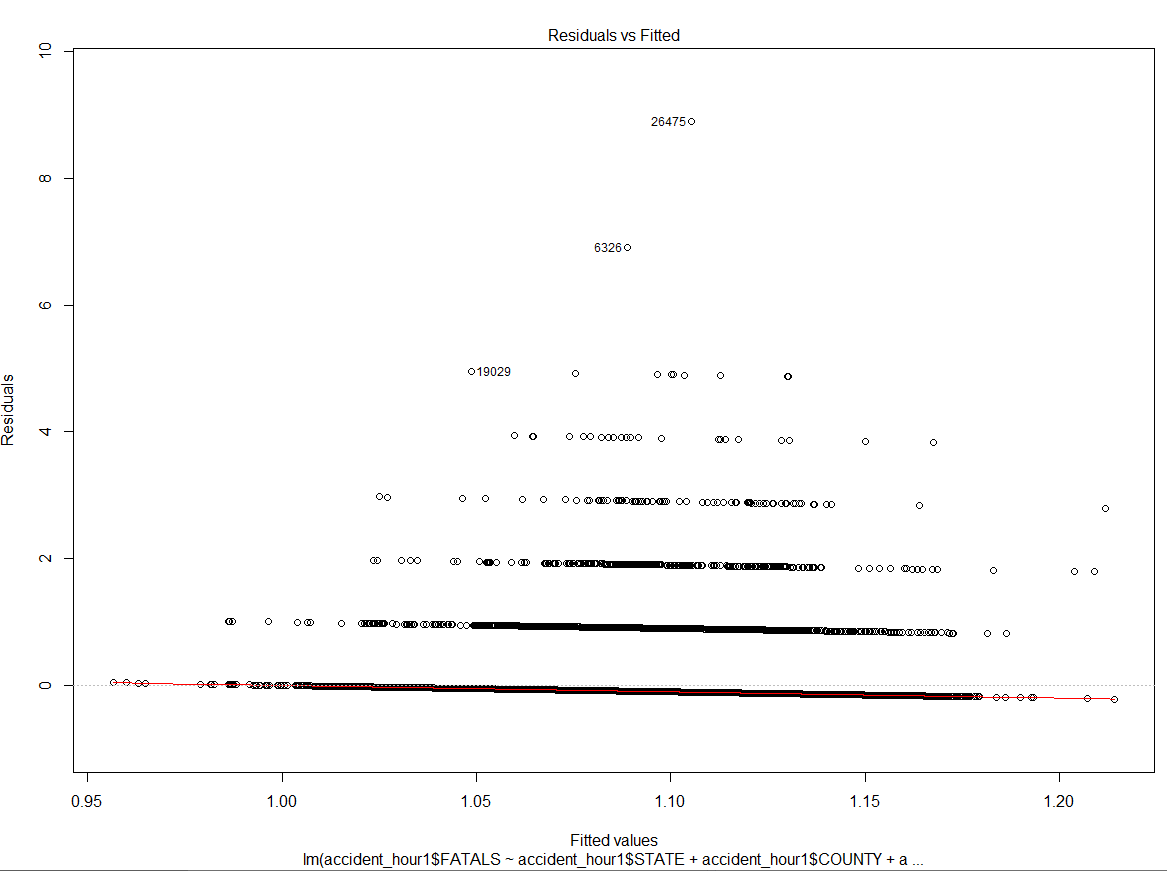
Residual standard error: 0.3657 on 31909 degrees of freedom

Multiple R-squared: 0.004924, Adjusted R-squared: 0.004705

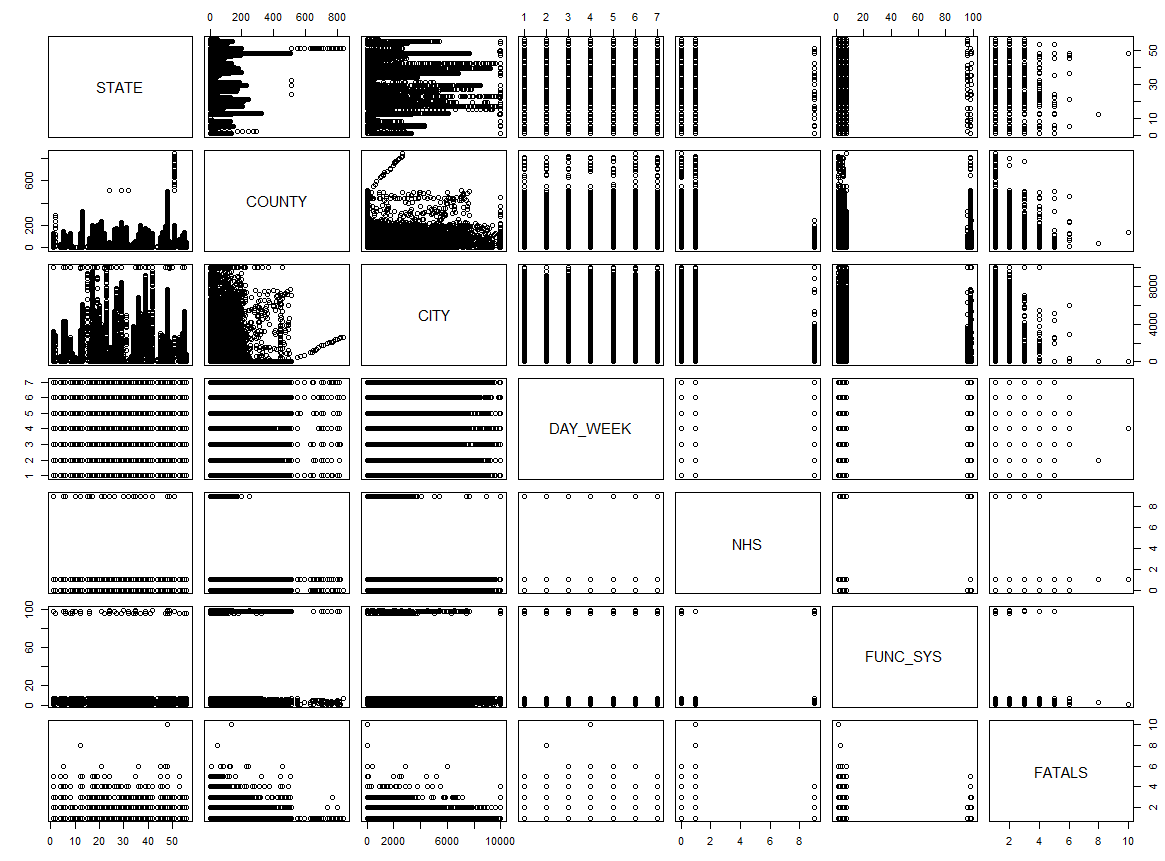
F-statistic: 22.55 on 7 and 31909 DF, p-value: < 2.2e-16

Although County, City, NHS, FUNC\_SYS, and DRUNK\_DR were statistically significant, the multiple R-squared value of the model is 0.004924, suggesting only 0.5% of the fatalities can be accounted for using these variables.

This is also shown when plotting the fitted values against the residuals:



This can be further confirmed when plotting the independent variables against one another:



Since some of the variables are categorical, indicator variables must be created to represent the levels in a given category. By doing this, the dependency of the fatalities would be more accurately represented as a function of the independent variables.