APPENDIX D

Regression Analysis for Safety, Traffic & Regulatory Components

Regression analysis was used to develop a model to estimate the safety, traffic and regulatory needs. As discussed in Chapter 4, multiple models were examined before the final model was selected.

The final model considered total replacement cost as the response variable and total miles, agency type and climate type as predictors. The variables agency type and climatic region are indicator variables and do not have a natural scale or measurement. They were used to group the data and account for variations not explained with quantitative variables.

The indicator variables used in this model are described below.

Agency Type:

- Urban: Urban miles ≥ 75% of total miles
- Rural: Urban miles ≤ 25% of total miles
- Combined: Urban miles between 26% and 74% of total miles

Climatic Region:

- Central: Central Coast, South Coast, Inland Valley
- Coast: North Coast, Low Mountain, South Mountain
- Mountain: High Mountain, High Desert
- Desert: Desert
- Mixed: Any combination of regions

The climatic regions were based on Caltrans specifications for PG binder grade selection and are shown in Figure D.1.

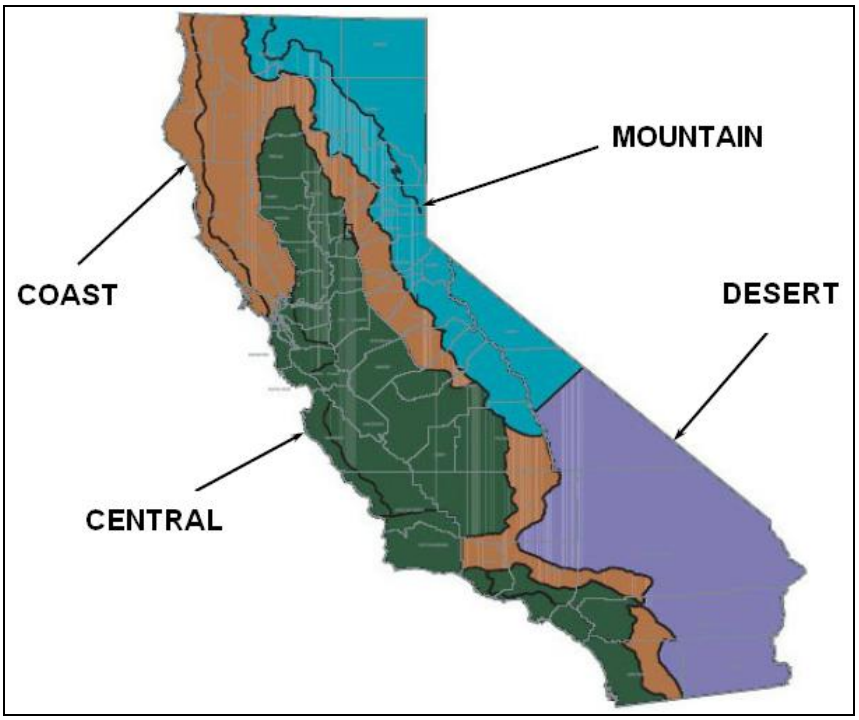

Figure D.1. Caltrans Performance Grade Binder Map

Indicator variables have values of 0 and 1 to identify the different types described above. For example, in the regression the variable agency type was defined as follows:

Type_Urban	=	1 if the agency is urban 0 otherwise
Type_Rural	=	1 if the agency is rural 0 otherwise
Type_Combined	=	1 if the agency is combined 0 otherwise

Once the variables were defined, the next step was to perform a multiple regression between the response and all the possible predictors. The output of the regression provides several parameters that were used to evaluate the model:

- **Analysis of Variance:** This approach was used to test the significance of the regression. If the p-value from the analysis of variance is < 0.05, it indicated that there was a linear relationship between the response and at least one of the predictors (at a 95% confidence level).

- **p-values for individual coefficients:** these values indicated the significance of each predictor within the model. p-values < 0.05 indicate that the predictor was highly significant at 95% confidence level.

- **Variance Inflation Factors (VIF):** These values were used to identify multicollinearity (strong correlation among the predictors), which can dramatically impact the ability to estimate regression coefficients. VIFs larger than 10 imply serious problems with multicollinearity.

- **R^2 and adjusted R^2**: R^2 indicates the proportion of variation explained by the predictors. Values of R^2 close to 1 imply that most of the variability in the response was explained by the regression model. The adjusted R^2 penalizes the addition of variables that were not significant to the model and was useful in evaluating and comparing candidate regression models.

In addition, the adequacy of the model was checked to ensure that the following assumptions were met:

- The relationship between the response and the predictors was linear.
- The error term had constant variance (was homogeneous)
- The errors were normally distributed.

Figure D.2 is the output from the initial regression. The p-value from the analysis of variance was < 0.05, which indicated that there was a relationship between at least one of the predictors and the total cost. VIFs < 10 indicate that there were no multicollinearity problems. R^2 = 52.3% indicate that there was about 48% of the variability not explained by the model.

```
The regression equation is
TOTAL COST = - 2.45E+09 + 2205308 TOTAL MILES - 8.67E+08 TYPE_RURAL
             + 1.21E+09 TYPE_URBAN + 1.33E+09 CLIMATE_CENTRAL
             + 1.23E+09 CLIMATE_COAST


Predictor               Coef    SE Coef      T      P    VIF
Constant         -2447667582  787650028  -3.11  0.003
TOTAL MILES          2205308     315946   6.98  0.000  2.385
TYPE_RURAL         -867135945  641001247  -1.35  0.182  1.494
TYPE_URBAN         1209008158  468666129   2.58  0.013  2.545
CLIMATE_CENTRAL    1331953147  659775232   2.02  0.049  7.799
CLIMATE_COAST      1233703077  704556212   1.75  0.086  7.734


S = 728821732   R-Sq = 52.3%   R-Sq(adj) = 47.7%


Analysis of Variance

Source          DF           SS           MS      F      P
Regression       5  3.03235E+19  6.06470E+18  11.42  0.000
Residual Error  52  2.76214E+19  5.31181E+17
Total           57  5.79449E+19
```

Figure D.2 Initial Regression Output

Model Adequacy Checking

Several basic assumptions were made when building the initial model:

- The relationship between the response and the predictors was linear.
- The error term was homogeneous (constant variance).
- The errors were normally distributed.

It was necessary to examine the adequacy of the proposed model because any violations of the assumptions above may yield an unstable model. The residual analysis method was used in this study. Residuals are a measure of the variability in the observations not explained by the regression model and can identify departures from the model assumptions. Studentized residuals are adjusted residuals with constant variance that provide a better scale. The following are graphical methods used to check the model assumptions:

- **Linearity:** Plot residuals versus fitted values. If a curve band or a non-linear pattern showed up, then either polynomial terms or a transformation should be considered (Figure D.3).

- **Constant Variance:** Plot studentized residuals versus fitted values. If scatter increased with fitted values, the errors have non-constant variance (Figure D.4).

- **Normality:** Create a normal probability plot by plotting the ordered studentized residuals versus the expected order statistics from a standard normal distribution. If the resulting plot produces points close to a straight line then the data are consistent with that from a normal distribution (Figure D.5).

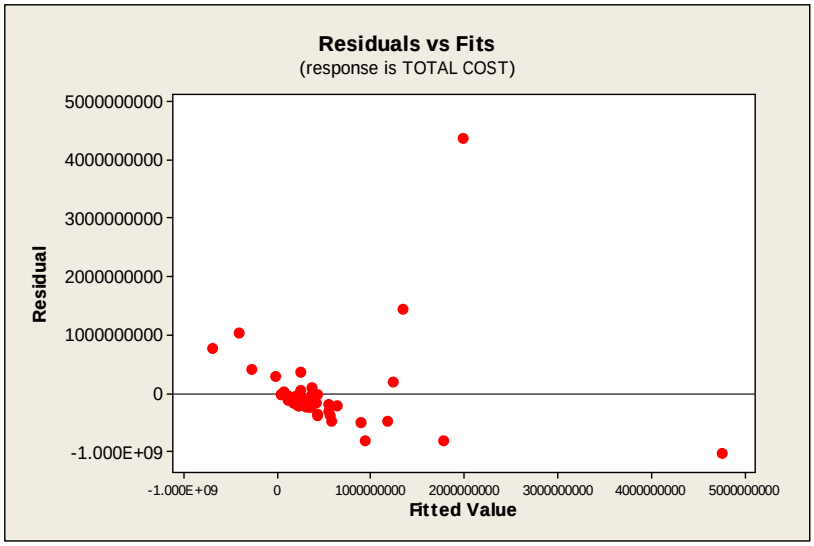

Figure D.3. Residual Plot

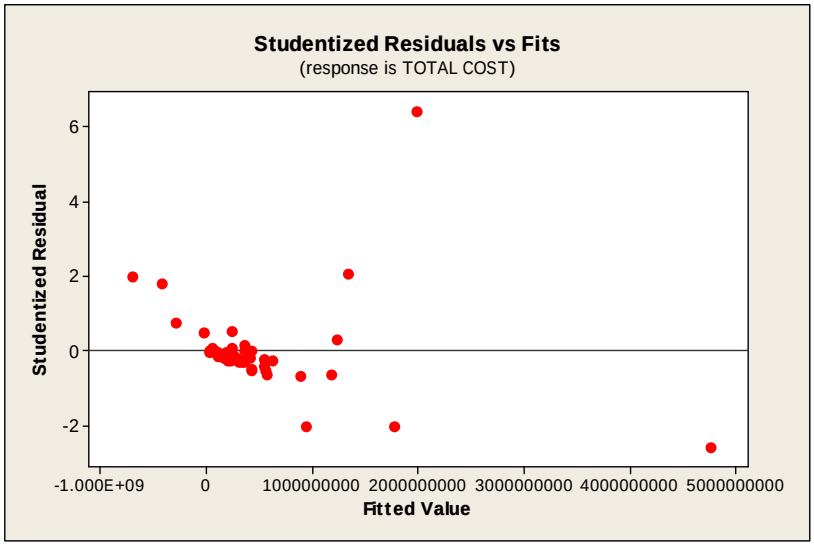

Figure D.4 Studentized Residual Plot

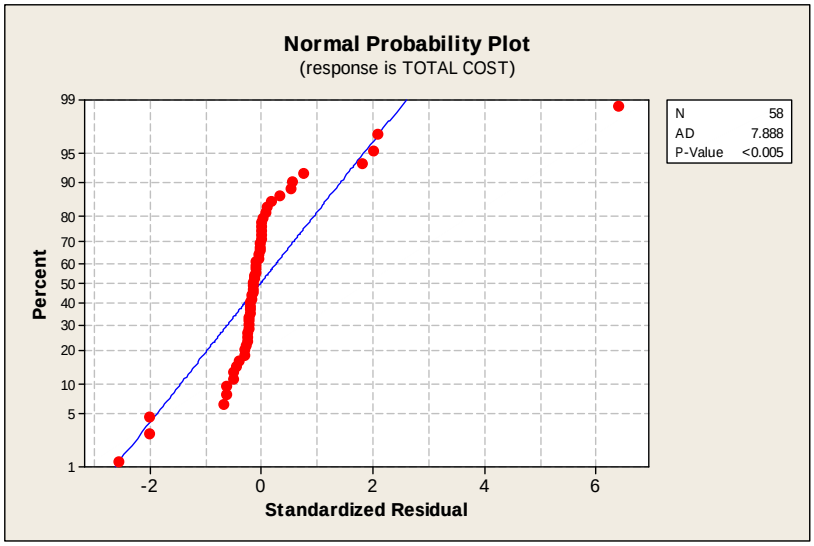

Figure D.5 Normal Probability Plot

From Figures D.4 through D.5, it can be observed that the model assumptions of constant variance and normality were violated.

Detection of Outliers

Outliers are data points which are not typical of the rest of the data. If the studentized residual fell outside the interval -2 to 2, the point was considered an outlier. If outliers were detected, they were thoroughly investigated before any actions were taken. The following outliers were detected:

Table D.1. Outliers Detected

County	Agency	Studentized Residual
Orange	Huntington Beach	2.07
San Diego	San Diego	-2.57
San Diego	San Diego County	-2.01
San Francisco	San Francisco	6.40
San Luis Obispo	San Luis Obispo County	-2.01
Shasta	Shasta County	2.01

No action was taken because these data points correspond to large agencies that should be considered in the analysis.

Leverage and Influence Points

Leverage and influence points have considerable influence on the fitted model. A leverage point is a point whose x-value is distant from the other x-values. It does not affect the estimate of the regression coefficients but will have a significant impact on the model summary statistics such as R^2. Influence points have both x and y-values that are distant from the other data points and have noticeable impact on the model coefficients.

The following unusual observations were identified:

Table D.2 Leverage and Influence Points

County	Agency	Leverage Point	Influence Point
Marin	Marin County	X	
San Diego	San Diego	X	X
San Diego	San Diego County	X	X
San Luis Obispo	San Luis Obispo County	X	
San Francisco	San Francisco		X
San Luis Obispo	San Luis Obispo County		X
San Mateo	San Mateo County	X	
Shasta	Shasta County	X	X

Once again, no action was taken because there is no reason to doubt the validity of the data.

Correction of Model Inadequacies

Model inadequacies can sometimes be corrected through data transformation. A log transformation was applied to the response in order to stabilize the variance and normalize the distribution of the errors. The output and residual analysis from the transformed model are shown in Figures D.6 through D.9.

```
The regression equation is
LOG COST = 17.0 + 0.00216 TOTAL MILES - 1.74 TYPE_RURAL + 0.442 TYPE_URBAN
           + 1.11 CLIMATE_CENTRAL + 0.541 CLIMATE_COAST


Predictor             Coef    SE Coef      T      P    VIF
Constant            16.980      1.053  16.13  0.000
TOTAL MILES      0.0021573  0.0004223   5.11  0.000  2.385
TYPE_RURAL         -1.7438     0.8567  -2.04  0.047  1.494
TYPE_URBAN          0.4422     0.6264   0.71  0.483  2.545
CLIMATE_CENTRAL     1.1131     0.8818   1.26  0.212  7.799
CLIMATE_COAST       0.5415     0.9416   0.58  0.568  7.734


S = 0.974061   R-Sq = 46.0%   R-Sq(adj) = 40.9%


Analysis of Variance

Source          DF       SS      MS     F      P
Regression       5  42.0968  8.4194  8.87  0.000
Residual Error  52  49.3373  0.9488
Total           57  91.4342
```

Figure D.6 Transformed Regression Output

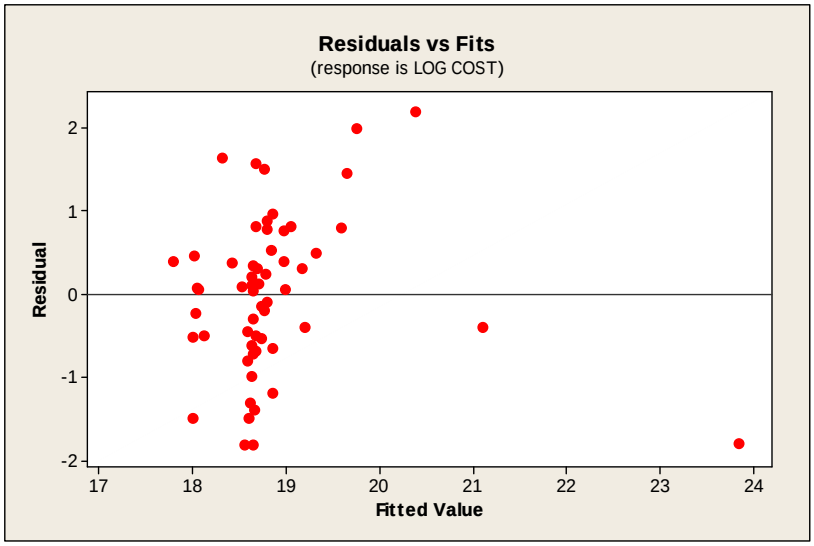

Figure D.7 Residual Plot

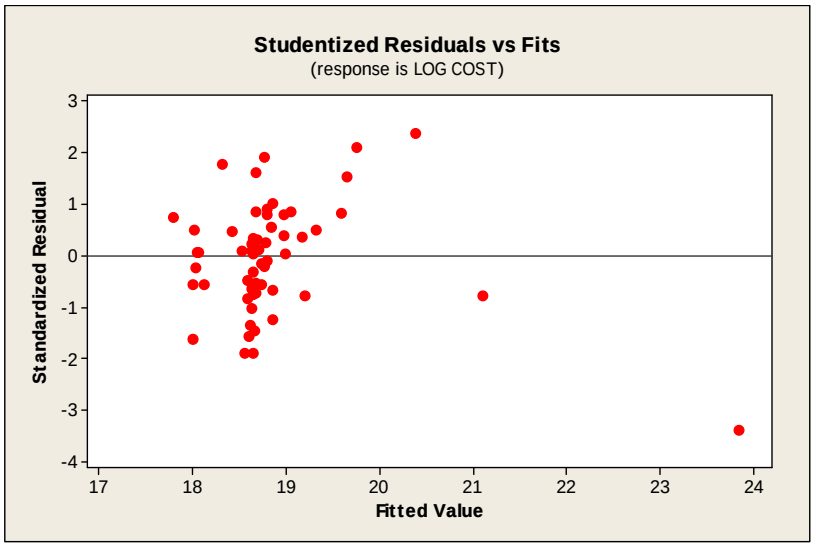

Figure D.8 Studentized Residual Plot

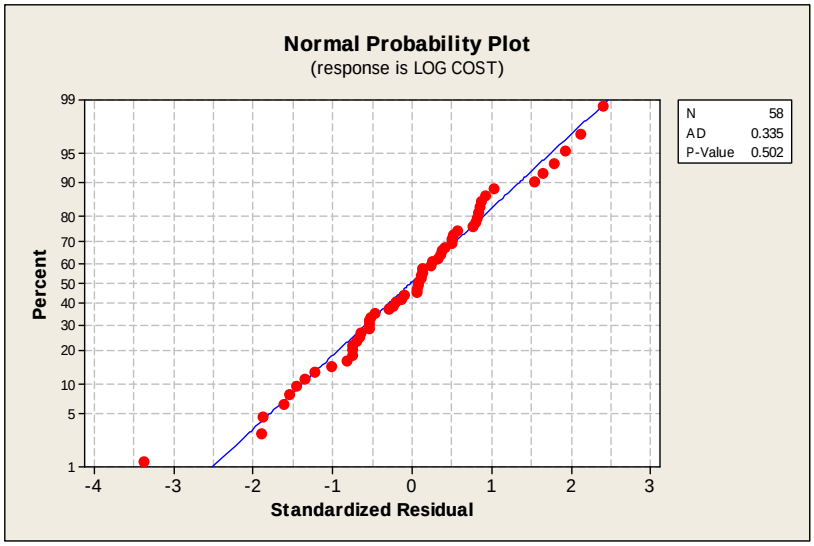

Figure D.9 Normal Probability Plot

The residual plots show that the model inadequacies have been corrected with the transformation.

Variable Selection

Variable selection is a technique used to ensure that all the predictors in the model are significant. Using stepwise regression methods, it was determined that only the following predictors contribute to the model:

- Total Miles
- Type_Rural
- Climate_Central

Figure D.10 shows the output of the reduced model.

```
The regression equation is
LOG COST = 17.9 + 0.00189 TOTAL MILES - 2.09 TYPE_RURAL + 0.682
CLIMATE_CENTRAL


Predictor             Coef    SE Coef      T      P    VIF
Constant           17.8872     0.2948  60.67  0.000
TOTAL MILES      0.0018856  0.0002957   6.38  0.000  1.194
TYPE_RURAL         -2.0947     0.7616  -2.75  0.008  1.206
CLIMATE_CENTRAL     0.6818     0.3158   2.16  0.035  1.021

S = 0.963894   R-Sq = 45.1%   R-Sq(adj) = 42.1%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       3  41.263  13.754  14.80  0.000
Residual Error  54  50.171   0.929
Total           57  91.434
```

Figure D.10 Reduced Regression Output

Final Model

The final model that met all these requirements was as follows:

log Cost = 17.9 + 0.00189 Total Miles – 2.09 Type_Rural + 0.682 Climate_Central

It should be noted that:

- If the agency type was "Urban" or "Combined" or if the climatic region is other than "Central" the indicator variables will have a value of zero and the model will depend only on total centerline miles.
- "log" refers to the natural logarithm

Table D.3 below is an ***example*** of the estimation of the safety, traffic and regulatory needs for an analysis period of 25 years and a total replacement cost of $1.0 billion.

Table D.3 Example of 25 Year Safety, Traffic & Regulatory Needs Calculations

Asset	% of Total Repl. Cost (1)	Replacement Cost (2)	Service Life (3)	Annual Needs (4)	25 Yr Needs (5)
Storm Drain	27.0	269,594,241	50	5,391,885	5,391,885
Curb & Gutter	26.1	260,972,222	35	7,456,349	7,456,349
Sidewalk	28.5	284,676,623	35	8,133,618	8,133,618
Curb Ramps	2.75	27,506,916	35	785,912	785,912
Traffic Signals	7.09	70,926,984	40	1,773,175	1,773,175
Street Lights	4.15	41,486,571	30	1,382,886	1,382,886
Sound/Retaining Walls	3.38	33,768,503	30	1,125,617	1,125,617
Traffic Signs	1.11	11,067,939	10	1,106,794	1,106,794
Total	**100**	**1,000,000,000**	**--**	**27,156,235**	**678,905,868**

Column (2) = $1.0 billion x Column (1)
Column (4) = Column (2) / Column (3)
Column (5) = Column (4) x 25 years