APPENDIX D

User Guide for CMF Regression Software

Contents

Introduction

The CMF Regression Software is a Microsoft Excel™ tool developed under *NCHRP Project 17-63: Guidance for the Development and Application of Crash Modification Factors* that can be used to conduct a statistical evaluation of crash modification factors (CMFs) for a common treatment or change in site characteristic. The evaluation can consist of: (1) computing the overall average CMF, (2) computing CMFs by crash type or severity category using reported aggregate CMFs, or (3) computing CMFs as a function of site characteristics.

The methods used in this software are based on the premise that CMF values (and their respective standard errors) for two or more locations are available from the literature and that the values are unbiased. They are also based on the premise that more detailed data (e.g., crash counts) are not available for the collective set of locations. There are more robust procedures available for computing an overall average CMF when detailed data (e.g., crash data) are available for the collective set of similar locations (e.g., Hauer, 1997). These procedures should be used to compute the overall average CMF when detailed data are available.

The evaluation is based on a weighted statistical analyses, where the reported standard error of the CMF is used to compute its weight. Statistical tests are implemented to assist the analyst in determining whether the CMFs are more appropriately described using an overall average value or a function of site characteristics. Weighted regression is used to quantify the regression coefficients.

The three evaluation types supported by the tool consist of: (1) computing the overall average CMF, (2) computing CMFs by crash type or severity category using reported aggregate CMFs, or (3) computing CMFs as a function of site characteristics. The key pieces of information needed to use the tool for each evaluation type are listed in Table D1.

Table D1. Evaluation types supported by CMF Regression tool.

Evaluation Type	**Data Needed**			
	CMF value	**CMF Standard Deviation**	**Crash Distribution**	**Site Characteristics**
1. Compute overall average CMF	Required	Required	--	--
2. Compute CMFs by crash type or severity (i.e., using aggregate CMFs) and, optionally, by site characteristics	Required	Required	Required	Optional
3. Compute CMFs as a function of only site characteristics	Required	Required	--	Required

Note: "—" – data not needed.

The CMFs used in the tool can be obtained from any source. Sources include the *Highway Safety Manual* (HSM), the CMF Clearinghouse, or one or more research publications that each document the development of a CMF for a common treatment. The first two sources (i.e., HSM and Clearinghouse) do not provide the crash distribution data (e.g., the distribution of crashes by severity type) so the conduct of evaluation type 2 will require the analyst to obtain the desired information from one or more CMF research documents.

Software Requirements

This spreadsheet tool was designed to work on Microsoft Excel™ 2007 and later.

The Solver add-in must be installed for the tool to work properly.

- For EXCEL 2007, select the Tools menu, choose Add-Ins.
- For EXCEL 2013, select Developer, choose Add-Ins.

If Solver is listed, the check its check box. If Solver is not listed, then click Browse, navigate to the folder where Solver.xlam is located, click on Solver.xlam, and click OK to close the Browse box. Click OK to close the Add-In box. If Solver was just installed for the first time in Excel, it will need to be used at least once manually before it can be used by this spreadsheet. That is, Solver is not "opened" for use by macros until it has been used once manually.

User Guide

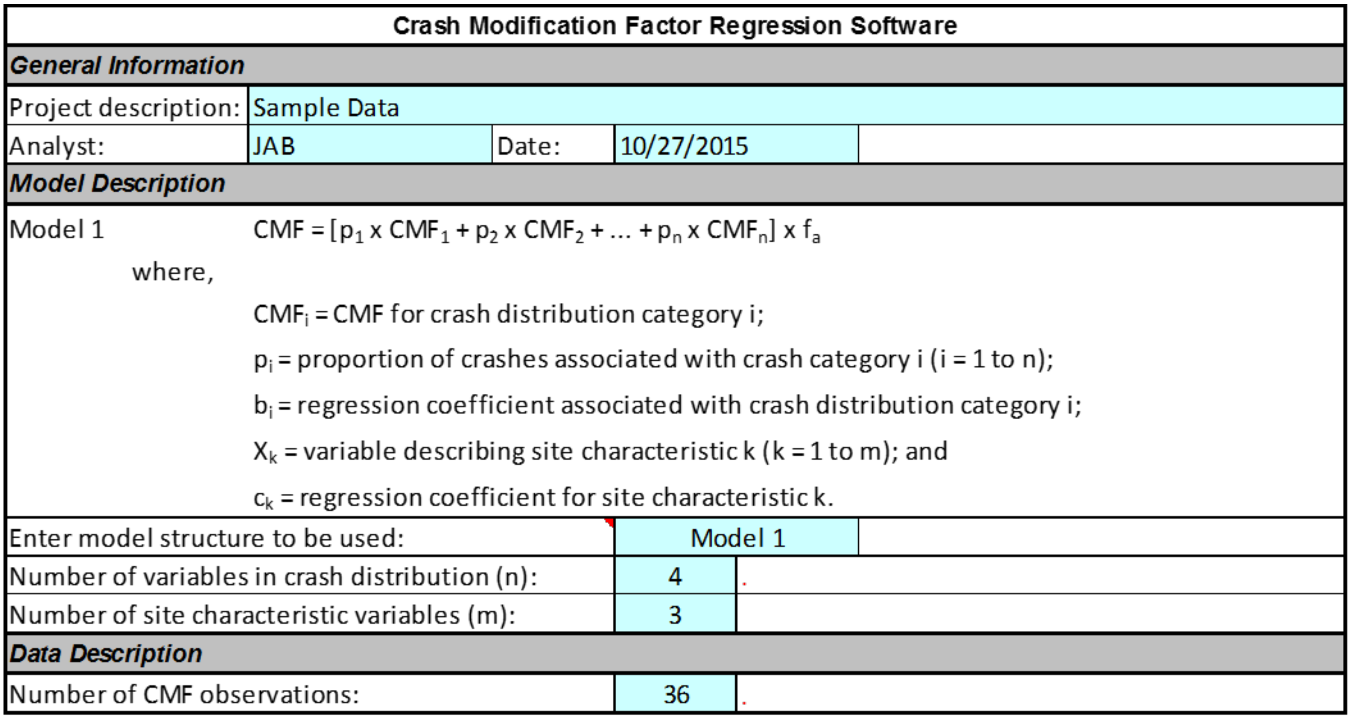
This software includes the following worksheets:

● Welcome   Includes a foreword, acknowledgements, and disclaimer.
● Introduction   Brief overview of the software and guidelines for its use.
● Set Up   Input data to describe the model structure and database.
● Main   Input CMF data and start calculations.

Each worksheet is designed in a consistent manner and their use is similar. The cells with a blue background are for user input. Most of the other cells are locked to prevent inadvertent changes to cell contents. The structure and input areas of each worksheet are described in the following sections.

Set Up Worksheet

The Set Up worksheet allows the analyst to enter data describing the project, model to be used, and number of CMFs to be evaluated. Figure D1 shows a screenshot of the Set Up worksheet.



Crash Modification Factor Regression Software

General Information

Project description:	Sample Data		
Analyst:	JAB	Date:	10/27/2015

Model Description

Model 1 $CMF = [p_1 \times CMF_1 + p_2 \times CMF_2 + \ldots + p_n \times CMF_n] \times f_a$

where,

CMF_i = CMF for crash distribution category i;

p_i = proportion of crashes associated with crash category i (i = 1 to n);

b_i = regression coefficient associated with crash distribution category i;

X_k = variable describing site characteristic k (k = 1 to m); and

c_k = regression coefficient for site characteristic k.

Enter model structure to be used:	Model 1
Number of variables in crash distribution (n):	4
Number of site characteristic variables (m):	3

Data Description

Number of CMF observations:	36

Figure D1. Set Up Worksheet

General Information

The analyst can enter any descriptive information about the location being analyzed. Use the blue cells to enter this data.

Model Description

The model structure to be used is entered. Two choices are currently supported: "Model 1" and "User Defined." The structure of Model 1 is described in equation form in this section of the worksheet. If selected, the software will enter the equation in the column headed "Predicted CMF" in the Main worksheet. It will use this equation to compute the predicted CMF for each crash distribution category.

If "User Defined" is selected, the analyst will need to enter the equation in the column headed "Predicted CMF" in the Main worksheet. The equation will need to be entered using Excel formula protocols. It will need to be copied to each row for which there is an observation. The cells in which an equation is to be entered will have a grey background.

The number of variables used to describe the crash distribution is entered in the blue cell provided. This number must have a value of one or larger. The crash distribution variables can describe the crash severity distribution, the crash type distribution, or a combination of both crash severity and type.

The number of site-characteristic variables is entered in the blue cell provided. This number must have a value of zero or larger. The analyst can include any site characteristics that he/she believes may influence the value of the CMF. Example site characteristics include: area type (i.e, urban/rural), traffic volume, lane width, etc.

If a site characteristic has "n" levels, then it can be represented as a series of "n-1" indicator variables. For example, if the variable describes area type and it has 2 levels (i.e., urban or rural), then one (=2-1) site characteristic variable (e.g., urban) is represented in the data. This variable has a value of 1.0 if the site is urban and 0.0 if the site is not urban.

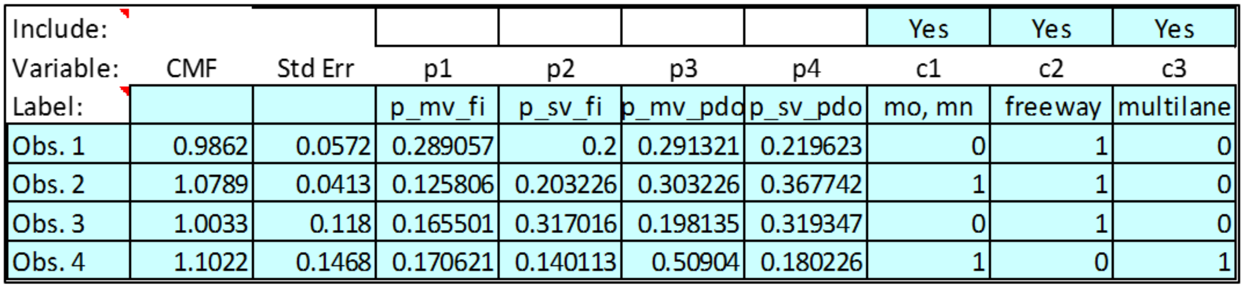
A site characteristic variable can be entered into the Main worksheet using its original units, or it can be transformed to alternative units if desired. For example, if an analysis of the data indicates that the reported CMFs vary with the natural log of the AADT, then the natural log of the AADT should be entered in the Main worksheet as a site characteristic variable.

Data Description

The number of CMF observations is entered in the blue cell provided. The minimum number of observations must exceed by 2 the sum of the crash distribution and site characteristic variables. The spreadsheet can be used to evaluate up to 500 observations.

Main Worksheet

The Main worksheet is where the analyst enters the CMF information and begins the calculations. Figure D2 shows a portion of the Main worksheet.



Include:							Yes	Yes	Yes
Variable:	CMF	Std Err	p1	p2	p3	p4	c1	c2	c3
Label:			p_mv_fi	p_sv_fi	p_mv_pdo	p_sv_pdo	mo, mn	freeway	multilane
Obs. 1	0.9862	0.0572	0.289057	0.2	0.291321	0.219623	0	1	0
Obs. 2	1.0789	0.0413	0.125806	0.203226	0.303226	0.367742	1	1	0
Obs. 3	1.0033	0.118	0.165501	0.317016	0.198135	0.319347	0	1	0
Obs. 4	1.1022	0.1468	0.170621	0.140113	0.50904	0.180226	1	0	1

Figure D2. Portion of Main worksheet

The reported CMF data are entered in the blue cells shown in the figure above. Sample data are shown for four observations (note: these data can be cleared by clicking on the Clear Data button, described below). The first column is used to enter text to describe each observation (e.g., "Obs 1", "Obs 2", etc.). The first blue row is used to enter text to describe each variable. For example, the data in the column headed "p1" are labeled "p_mv_fi" indicating the column contains data describing the proportion of multiple-vehicle fatal-and-injury crashes.

The reported CMF values are entered in the blue cells in the column headed "CMF". The standard error associated with each reported CMF is entered in the column headed "Std Err".

If data are to be entered by copying data from another spreadsheet into the blue cells of the Main worksheet, then paste by "value" only (i.e., Ctrl-Alt-V, Value, OK). Do not use a simple paste, or similar, because it will delete the cell background color and borders. The software relies on the correct cell background color to work properly.

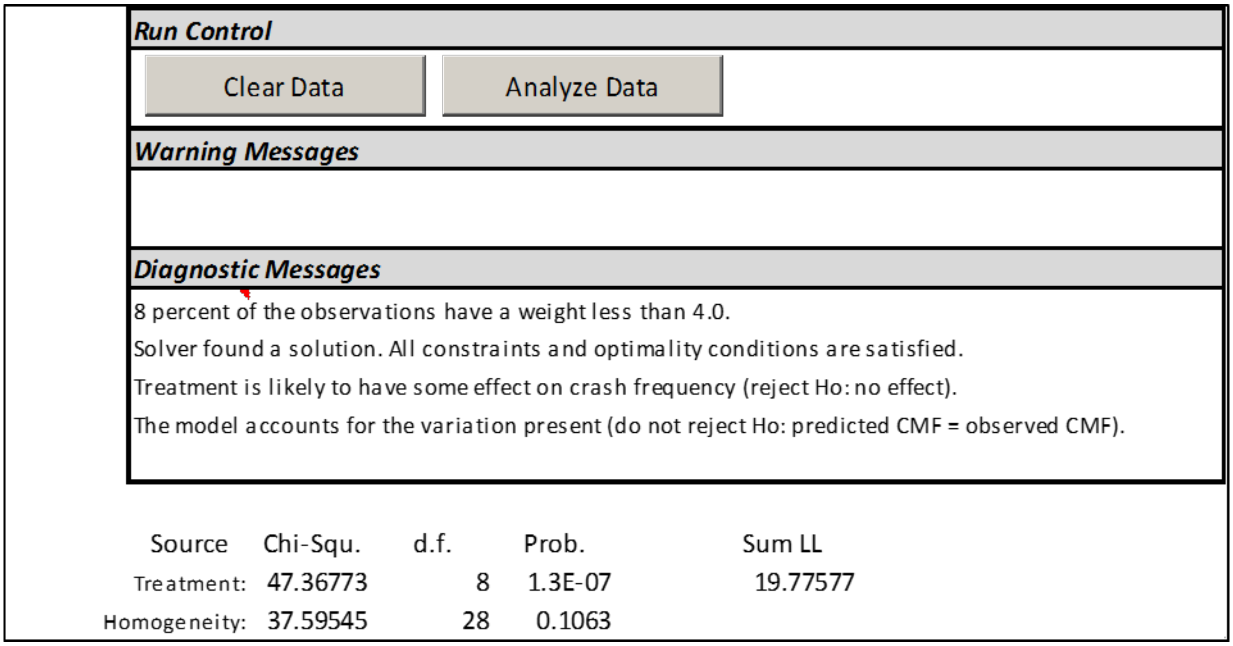
Crash Distribution Variables

Figure D2 shows four columns headed "p1", "p2", "p3", and "p4". These columns correspond to the four crash distribution variables that were indicated in the data entry of the Set Up worksheet. In general, the number of columns headed "p1", ..., "pn" will be equal to the number of crash distribution variables (n) indicated in the data entry of the Set Up worksheet. For a given observation (i.e., row), the crash distribution proportions should add to 1.000.

Site Characteristic Variables

Figure D2 shows three columns headed "c1", "c2", and "c3". These columns correspond to the three site characteristic variables indicated in the data entry of the Set Up worksheet. In general, the number of columns headed "c1", ..., "cm" will be equal to the number of site characteristic variables (m) indicated in the data entry of the Set Up worksheet.

The top row of Figure D2 is labeled "Include:". The analyst can optionally include or exclude site characteristic variables from the analysis. To include a variable, the analyst will need to enter "Yes" in the blue cell in the associated column. To exclude a variable, the analyst will need to enter "No." If a variable is excluded, the cell background and text color will be changed to white. The data will NOT be deleted but it will be excluded from the analysis. To restore the data in the analysis, change a "No" to a "Yes".



Run Control

Warning Messages

Diagnostic Messages

8 percent of the observations have a weight less than 4.0.

Solver found a solution. All constraints and optimality conditions are satisfied.

Treatment is likely to have some effect on crash frequency (reject Ho: no effect).

The model accounts for the variation present (do not reject Ho: predicted CMF = observed CMF).

Source	Chi-Squ.	d.f.	Prob.	Sum LL
Treatment:	47.36773	8	1.3E-07	19.77577
Homogeneity:	37.59545	28	0.1063	

Figure D3. Run Control portion of Main worksheet

Run Control

Two grey buttons are provided to control the evaluation. These buttons are shown in the figure above. The analyst clicks on a button to effect the indicated operation. One button is labeled "Clear Data". This button is used to erase the data in the blue cells. It can be used to clear old data from the spreadsheet, before entering new data.

The button labeled "Analyze Data" is used after the data are entered in the blue cells provided in the Main worksheet. When this button is clicked, the regression analysis is conducted and the results are shown in rows 13 to 17. Various diagnostic messages are provided in rows 9 to 13.

Warning Messages

When the "Analyze Data" button is clicked, the data are checked to ensure the values are appropriate. In particular, the blue cells are checked to ensure that values are entered (i.e., no cells are blank). If a blank cell is found, then the analysis is stopped and a warning message provided in the section labeled "Warning Messages". An exception is the blue cells used to label the columns or rows. The label cells can be left blank.

The entered CMFs and associated standard errors are checked to ensure that there are no values that are not positive (i.e., zero or negative). The analysis is stopped and a warning message is provided if one or more values are not positive.

If the model structure is indicated as "User Defined" in the Set Up worksheet, then the analyst must enter the desired equation in the cells with grey background. If these equations are not entered, then the analysis is stopped and a warning message is provided.

Excel Solver is used to compute the regression coefficients. If Solver is unable to converge on a reliable set of coefficients then a message is provided and the analysis is stopped.

Diagnostic Messages

Several diagnostic messages are provided to help the analyst interpret the analysis results. These messages are provided in the section labeled "Diagnostic Messages" of the Main worksheet. Example messages are shown in Figure D3.

- The first message reports the percentage of observations with a weight less than 4.0. CMFs with a smaller weight are less reliable because they are based on a relatively small sample size (i.e., a small number of observed crashes). A majority (desirably all) of the CMFs should have a weight of 4.0 or more.

- Excel Solver is used to compute the regression coefficients. If Solver is able to converge on a reliable set of coefficients then a message is provided and the computed CMFs are displayed.

- The third diagnostic message describes the results from the Chi-Square test of the treatment effect. The null hypothesis is that the treatment does not have an effect on safety. The statistics associated with this test are shown in the second-to-last row of the figure above. A significance level (alpha) of 0.05 is used to determine whether to reject the null hypothesis.

- The fourth diagnostic message describes the results from the Chi-Square test of homogenity in the CMF values. The null hypothesis is that predicted CMF equals the observed CMF. The statistics associated with this test are shown in the last row of the figure above. A significance level (alpha) of 0.05 is used to determine whether to reject the null hypothesis.

- The standard error of the regression coefficients is computed using the Jacobian matrix of first derivatives of the regression model coefficients. If this matrix is not positive definite, then the Jacobian matrix cannot be used. In this situation, the standard errors are estimated using the Jackknife technique. The need to use the Jackknife technique is indicated by a fifth diagnostic message (if this message is not presented, then the Jacobian matrix is used to estimate the standard error).

Standardized Residual

Figure D4 is produced in the Main worksheet. It displays the standardized residual for each observation. A standardized residual larger than +3.0 or smaller than -3.0 may be an outlier (i.e., input data may have an error, or observation may from a site with unexplained attributes that are different from those of the other sites). The standardized residual (r) is computed as r = (observed CMF - predicted CMF)/standard error.

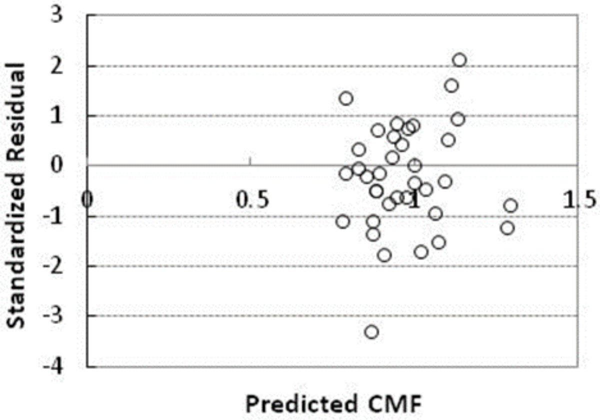


Figure D4. Plot of Standardized Residuals

Regression Coefficients

After the "Analyze Data" button is clicked, the best-fit regression coefficients are shown in the Main worksheet. The are shown in the second-to-last row of the figure above. These coefficients correspond to variables "p1", ... "pn", and "c1", ..., "cm", as defined in the Set Up worksheet. These coefficients are used in the equations in the column headed "Predicted CMF" to compute the predicted CMF for each observation.

The standard error of each regression coefficient is provided in the row below the corresponding regression coefficient. The standard error is used to estimate the variability in the coefficient values. Coefficients with a relatively large standard error are less confidently known than those coefficients with a small standard error. If a site characteristic variable is not known with the desired level of confidence, it can be excluded from the regression model by entering "No" in the "Include" row of the Main worksheet.

Average CMF

After the "Analyze Data" button is clicked, the best-fit regression coefficients are used to compute the average CMF for each crash distribution category. The average CMFs are shown in the second row of the figure above. If one or more site characteristic variables are included in the model, then the computed average CMF includes the effect of the site characteristics at the values indicated in the "Enter Value" row.

In the figure above, the analyst has entered "1", "1", and "0" for the three site characteristic variables "c1", "c2", and "c3", respectively. The variable "c1" is an indicator variable indicating the state whose data were used to quantify the CMF (= 1 for Minnesota or Missouri, 0 for Pennsylvania). The variable "c2" is an indicator variable for roadway class (= 1 for freeway, 0 for non-freeway). The variable "c3" is for roadway class (= 1 for multilane highway, 0 for non-multilane highway). Based on the three entered values shown, the four average CMFs listed in the figure above correspond to freeways in Minnesota or Missouri. The analyst can enter other values in the "Enter Values" row (and click on the Analyze Data button) to compute the average CMFs for other combinations of site characteristics.

In the figure above, the average CMF in the column headed "p1" is 1.004. It was noted in a previous figure that this column describes the proportion of multiple-vehicle fatal-and-injury crashes. So, the average CMF for multiple-vehicle fatal-and-injury crashes on freeways in Minnesota or Missouri is 1.004.

The standard error of the computed average CMF is also provided in the Main worksheet. This statistic is shown in the third row of the figure above. It is computed using the variance-covariance matrix of the regression model. The standard error of the CMF describes the distribution of the average CMF when the treatment is applied to similar sites. It is used to estimate the 95th percentile confidence interval of the average CMF. This confidence interval is shown in the fourth row of the figure above. The confidence interval is based on an assumed lognormal distribution of the average CMF.

In the figure above, the standard error of the average CMF in the column headed "p1" is 0.125. It was noted previously that this column describes the proportion of multiple-vehicle fatal-and-injury crashes. So, this statistic describes the standard error of the average CMF (i.e., 1.004) for freeways in Minnesota or Missouri. The 95th percentile confidence interval is 0.78 to 1.27.

If the analyst selects "User Defined" for the model form, then the average CMF and standard error cannot be computed because the derivatives of a user defined model form are unknown.

Application Process

The following steps describe the use of this spreadsheet to conduct a regression analysis of CMFs. The process is also illustrated in an example application later.

Step 1. Describe the Model and Database in the Set Up Worksheet

The analyst goes first to the Set Up worksheet to describe the model and the data.

Step 2. Enter the Data in the Main Worksheet

The analyst clicks on the "Clear Data" button to clear all data remnants from prior analyses. Then, the data are entered in the blue cells. These data include the reported CMFs and their standard errors, as well as the crash distribution data (if disaggregate CMFs are sought) and the site characteristics (if CMF functions are sought).

If the analyst has chosen "User Defined" model in the Set Up worksheet, then the equation for the model is entered in each of the cells with a grey background. The desired equation must use Excel formula protocols. This equation must then be copied down for all rows with observations.

Step 3. Enter Values of the Site Characteristic Variables

If site characteristic variables are used and if the model form is not User Defined, then the site characteristics for which the average CMF will be computed are entered in the blue cells on the "Enter Value" row. That is, if in the Set Up worksheet a pre-defined model is selected and it includes one or more site characteristic variables, then the analyst should enter values for the site characteristics for which an estimate of the average CMF is desired.

If no site characteristics are used or, if the model form is User Defined, then values do not need to be entered in the blue cells on the "Enter Value" row. This step can be skipped.

Step 4. Initiate Calculations

Click on the "Analyze Data" button to initiate the calculation process. Assess the validity of the observations using the standardized residuals. Correct or eliminate the observations as needed. Assess the variablity of the regression coefficients using the standard error of the regression coefficients. Include or exclude a site characteristic variable as needed. Consider the diagnostic messages. Determine if more (or less) site characteristic variables may be helpful.

Record the computed average CMFs for the specified site characteristic variables. Consider also the standard deviation of these CMFs.If the variability is large, then the average CMF value may not be sufficiently well known to represent the basis for reliable investment decisions. If the confidence interval includes 1.00, then there is a chance that the treatment may not improve safety at every site.

If it is desired to estimate CMF values for other combinations of site characteristics, then repeat Steps 3 and 4 using the new values for the site characteristic variables.

Overall Average CMF. To compute an overall average CMF value using this spreadsheet, in the Set Up worksheet enter "1" for the number of crash distribution variables; enter "0" for the number of site characteristic variables. In the Main worksheet, use the cells with blue background to enter "1" for each observation in the column headed "p1".

Example Application of CMF Regression Tool

The following section presents an example use of the CMF Regression tool to conduct evaluation type 1 and evaluation type 3. The CMF Clearinghouse was used as the source of CMFs for this application. The scenario posed for the example application is an analyst that desires to estimate the safety effect of two-way left-turn lane (TWLTL) installation on an existing two-lane roadway in the suburban area of a Colorado city. The AADT is 9,000 veh/d. The effect on total crash frequency is desired (inclusive of all crash types and severities).

The CMF Clearinghouse provides several CMFs to consider for this treatment. An examination of the CMFs provided indicates that the CMF value varies by area type (i.e., rural or urban). However, CMFs for suburban areas are not available and it is not clear whether the observed variation is so large as to suggest that treatment effect varies among states, or by area type. It is also noted that the AADT range associated with each CMF (with the exception of two CMFs) is not available in the Clearinghouse nor are other possible explanatory variables. The original research report would need to be consulted to identify the AADT and other site characteristics that may explain the CMF variation.

The values obtained from the Clearinghouse are listed in Table D2. There are eight CMFs (shown in Table D2). For this example, the eight CMF observations were developed for one study. They represent four states (i.e., Arkansas, California, Illinois, and North Carolina).

Table D2. CMFs for Install TWLTL on a two-lane road.

CMF ID	CMF Value	Standard Deviation	Area Type
2352	0.488	0.071	Rural
2353	0.962	0.083	Urban
2354	0.492	0.057	Rural
2355	1.028	0.134	Urban
2356	0.833	0.105	Rural
2357	0.906	0.100	Urban
2358	0.727	0.055	Rural
2359	1.05	0.088	Urban

The questions are: What is the one best estimate of the safety effect of a TWLTL? Does this value depend on area type, or the state in which the treatment was applied? If “no”, the overall average CMF should represent the best estimate of the CMF value for suburban areas in Colorado.

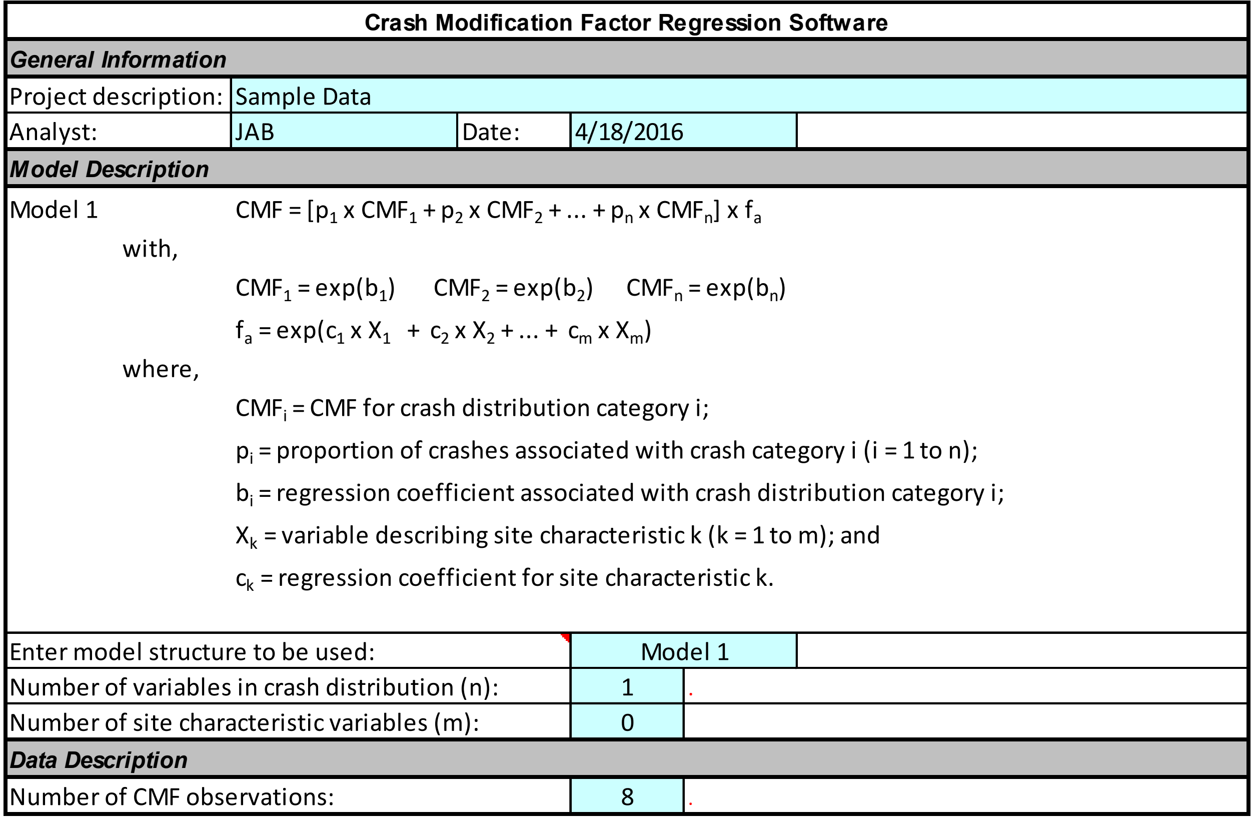
The Introduction worksheet in the CMF Regression Software tool describes the process for applying the tool as a sequence of steps. These steps are repeated in this section for the example application.

Step 1 Describe the Model and Database in the Set Up Worksheet

The Set Up worksheet is used for this step. The input data are shown in Figure D5.The model structure is selected to be “Model 1” because there is no reason to believe that an alternative form is appropriate.

As noted at the bottom of the Introduction worksheet, a “1” should be entered for the crash distribution variable input when an evaluation type 1 is conducted (i.e., compute an overall average CMF). More generally, this guidance holds whenever the crash distribution is not of interest to the analysis (i.e., for both evaluation type 1 and evaluation type 3). In this manner, the “p1” variable represents the intercept term in the regression model and there is one slope variable for each site characteristic (i.e., CMF = EXP[*intercept* + Σ*slope*$_i$ x *c*$_i$]).

The entry for the number of site characteristic variables is "0" because it is desired to answer the question of: Does this value depend on area type, state, or other variables?



Crash Modification Factor Regression Software

General Information

Project description:	Sample Data			
Analyst:	JAB	Date:	4/18/2016	

Model Description

Model 1 $CMF = [p_1 \text{ x } CMF_1 + p_2 \text{ x } CMF_2 + ... + p_n \text{ x } CMF_n] \text{ x } f_a$

with,

$CMF_1 = \exp(b_1)$ $CMF_2 = \exp(b_2)$ $CMF_n = \exp(b_n)$

$f_a = \exp(c_1 \text{ x } X_1 + c_2 \text{ x } X_2 + ... + c_m \text{ x } X_m)$

where,

CMF_i = CMF for crash distribution category i;

p_i = proportion of crashes associated with crash category i (i = 1 to n);

b_i = regression coefficient associated with crash distribution category i;

X_k = variable describing site characteristic k (k = 1 to m); and

c_k = regression coefficient for site characteristic k.

Enter model structure to be used:	Model 1
Number of variables in crash distribution (n):	1
Number of site characteristic variables (m):	0

Data Description

Number of CMF observations:	8

Figure D5. Set Up worksheet for example application.

Step 2. Enter the Data in the Main Worksheet

The analyst goes to the Main worksheet to complete this step. There, the first activity is to clear any data from a previous analysis. The data are cleared by clicking on the "Clear Data" button that is provided in the Run Control section of the worksheet. This button is shown on the left side of Figure D6.

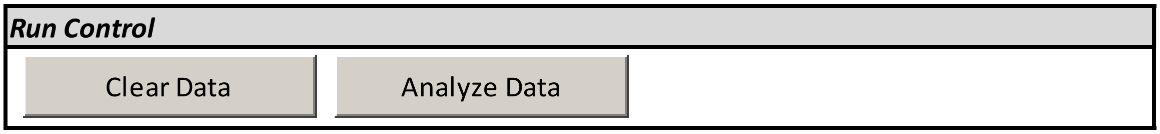
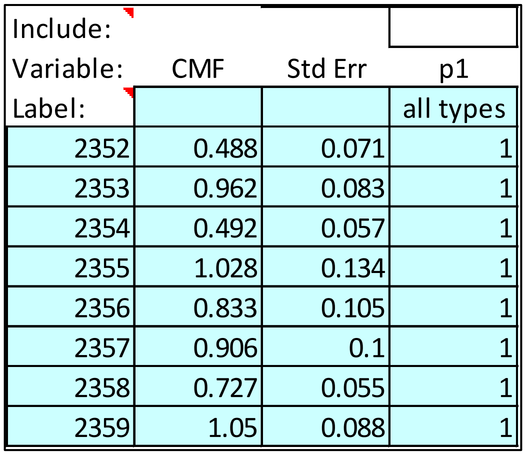


Figure D6. Run control buttons.

The data from Table D2 are entered into the blue cells of the Main worksheet. The "p1" column is provided for the crash distribution variable. A "1" is entered in this column because the crash distribution is not required for this application (i.e., an evaluation type 1 is being conducted). The entered data are shown in Figure D7.



Include:			
Variable:	CMF	Std Err	p1
Label:			all types
2352	0.488	0.071	1
2353	0.962	0.083	1
2354	0.492	0.057	1
2355	1.028	0.134	1
2356	0.833	0.105	1
2357	0.906	0.1	1
2358	0.727	0.055	1
2359	1.05	0.088	1

Figure D7. Entered data for example application.

Step 3. Enter Values of the Site Characteristic Variables

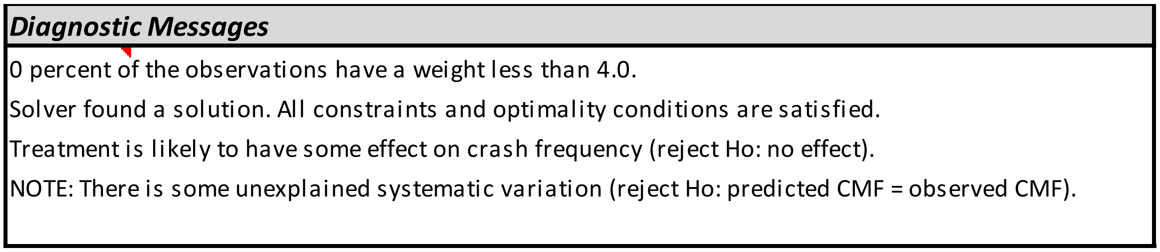
No site characteristic variables need to be examined to answer the question of interest, so this step is skipped.

Step 4. Initiate Calculations

The analyst then clicks on the Analyze Data button (as shown in Figure D6) to initiate the regression calculation process.

The standardized residuals are shown graphically at the top of the Main worksheet. Guidelines for using this figure to identify outliers are provided in the "Standardized Residual" section of the Introduction worksheet.

A series of diagnostic messages are provided in the Diagnostic Messages sections. Guidelines for interpreting these messages are provided in the "Diagnostics Messages" section of the Introduction worksheet. The diagnostic messages for this analysis are shown in Figure D8. The last message in the figure provides the results of the homogeneity test. It indicates that there is some unexplained systematic variation in the CMFs. The amount of variation is sufficiently large as to be statistically significant (i.e., *p* value is less than 0.05). As a result, the CMFs should not be combined. The safety effect of TWLTL installation on two-lane roads in suburban Colorado cannot be reliably estimated using the information in the Clearinghouse.



Diagnostic Messages

0 percent of the observations have a weight less than 4.0.

Solver found a solution. All constraints and optimality conditions are satisfied.

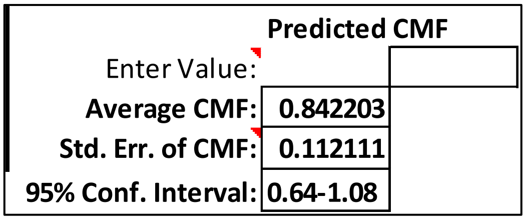
Treatment is likely to have some effect on crash frequency (reject Ho: no effect).

NOTE: There is some unexplained systematic variation (reject Ho: predicted CMF = observed CMF).

Figure D8. Diagnostic messages for example application.

At this point, the analyst's initial question is answered. It would not be appropriate to use the combined CMF to estimate the safety effect of TWLTL installation on a suburban road in Colorado. The analyst can now explore the cause for the observed variation in CMF values. One option is to assess whether the differences are explained by area type. A second option is to assess whether the differences are explained by state. The original research report could also be obtained to extract data for AADT or other site characteristics. These data would allow the exploration of whether the differences are due to differences in AADT or other site characteristics. The crash type distribution and crash severity distribution among locations could also be obtained from the original research report to determine whether the CMF differences are explained by differences in crash type or severity distribution.

Although it should not be used for this application, the overall average CMF is shown in Figure D9 as 0.842. The standard deviation of this CMF is 0.112 and the 95th percentile confidence interval is 0.64 to 1.08. Guidance for interpreting these statistics is provided in the section titled "Average CMF" of the Introduction worksheet.



	Predicted CMF	
Enter Value:		
Average CMF:	**0.842203**	
Std. Err. of CMF:	**0.112111**	
95% Conf. Interval:	**0.64-1.08**	

Figure D9. Computed CMF for TWLTL installation.

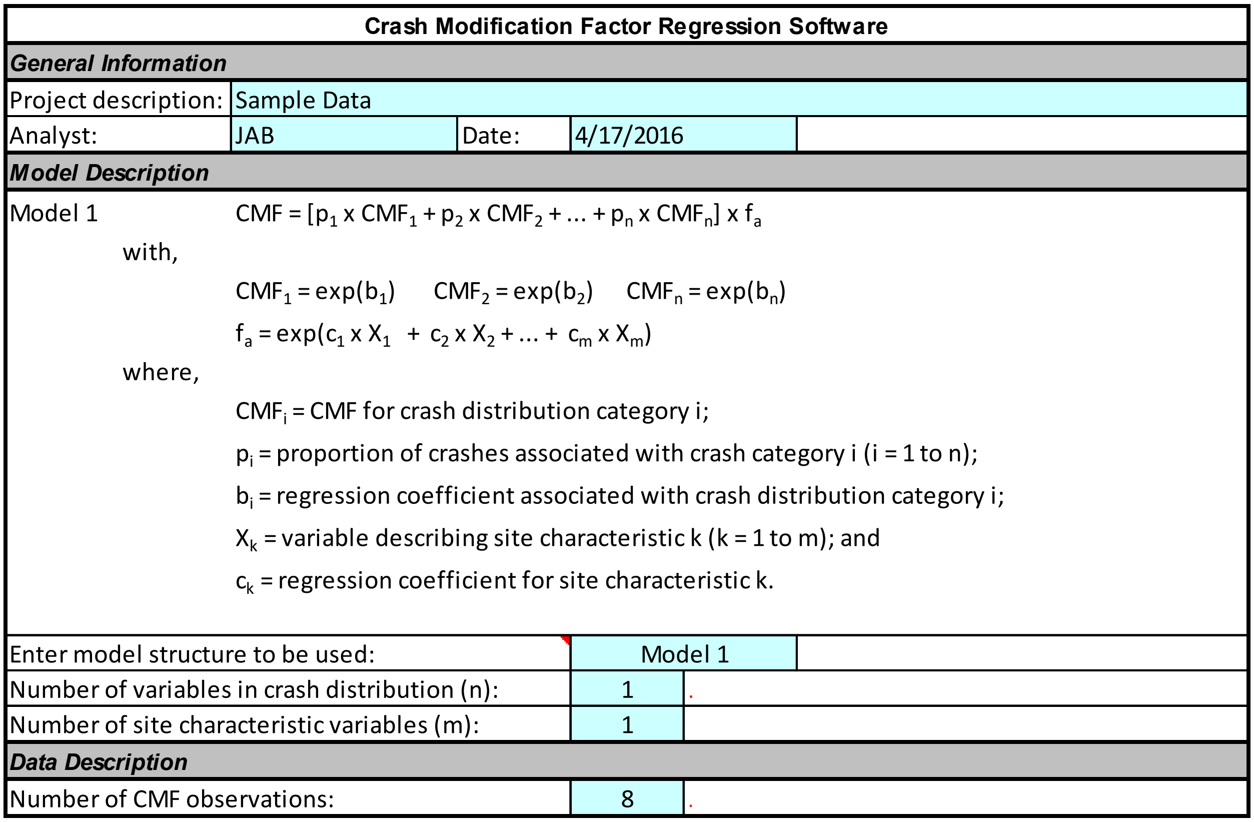
Example Application Continued

As a further demonstration of the software tool, the following paragraphs illustrate the exploration of the association between area type and CMF value.

Step 1 Describe the Model and Database in the Set Up Worksheet

The Set Up worksheet is used for this step. The input data are shown in Figure D10. The input information is the same as in the previous section, with the exception that a site characteristic variable is provided (i.e, this is an evaluation type 3).

The site characteristic of interest is area type. There are no other site characteristic variables of interest, so "1" is entered for the site characteristic variable input.



Crash Modification Factor Regression Software

General Information

Project description:	Sample Data			
Analyst:	JAB	Date:	4/17/2016	

Model Description

Model 1 $\quad CMF = [p_1 \text{ x } CMF_1 + p_2 \text{ x } CMF_2 + ... + p_n \text{ x } CMF_n] \text{ x } f_a$

with,

$CMF_1 = \exp(b_1) \quad CMF_2 = \exp(b_2) \quad CMF_n = \exp(b_n)$

$f_a = \exp(c_1 \text{ x } X_1 \; + c_2 \text{ x } X_2 + ... + \; c_m \text{ x } X_m)$

where,

CMF_i = CMF for crash distribution category i;

p_i = proportion of crashes associated with crash category i (i = 1 to n);

b_i = regression coefficient associated with crash distribution category i;

X_k = variable describing site characteristic k (k = 1 to m); and

c_k = regression coefficient for site characteristic k.

Enter model structure to be used:	Model 1	
Number of variables in crash distribution (n):	1	
Number of site characteristic variables (m):	1	

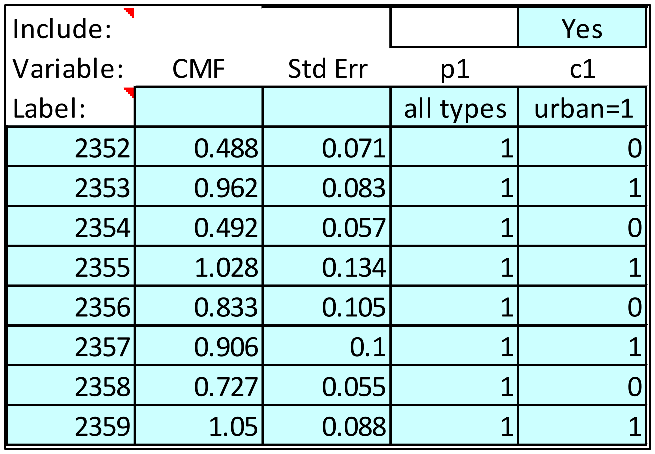
Data Description

Number of CMF observations:	8	

Figure D10. Set Up worksheet for example application.

Step 2. Enter the Data in the Main Worksheet

The analyst goes to the Main worksheet to complete this step. Most of the data were previously entered in Step 2 of the prior analysis. The entered data are shown in Figure D11. The new input data are shown in the last column. They are discussed in the next step.



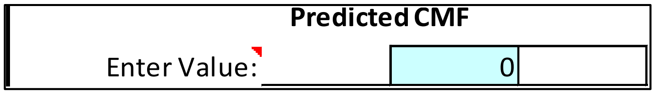
Include:				Yes
Variable:	CMF	Std Err	p1	c1
Label:			all types	urban=1
2352	0.488	0.071	1	0
2353	0.962	0.083	1	1
2354	0.492	0.057	1	0
2355	1.028	0.134	1	1
2356	0.833	0.105	1	0
2357	0.906	0.1	1	1
2358	0.727	0.055	1	0
2359	1.05	0.088	1	1

Figure D11. Entered area-type data for example application.

Step 3. Enter Values of the Site Characteristic Variables

The area type variable is categorical so it is converted to an indicator variable (i.e., numeric binary format) using "1" to represent urban and "0" to represent rural areas. These values are shown in the last column of Figure D11. Guidelines describing the conversion of categorical variables to indicator variables are provided in the Introduction worksheet, in the section titled Model Description.

The CMF for both rural and urban conditions is desired. The CMF for rural conditions is computed by entering a "0" in cell N11 (as shown below). The value "0" is used because this value was defined as representing rural conditions using the indicator variable described in the previous paragraph. For the same reason, if the CMF for urban areas is desired, then the value of "1" is entered. The calculation of the CMF for urban areas is discussed further in the next step.



	Predicted CMF	
Enter Value:	0	

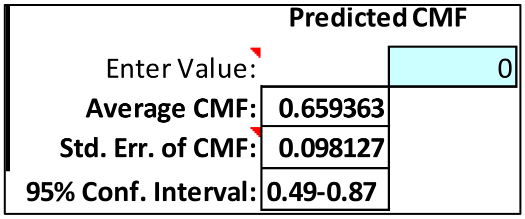
Step D4. Initiate Calculations

The analyst then clicks on the Analyze Data button (as shown in Figure D6) to initiate the regression calculation process.

A series of diagnostic messages are provided in the Diagnostic Messages sections. Guidelines for interpreting these messages are provided in the "Diagnostics Messages" section of the Introduction worksheet. The diagnostic messages for this analysis indicate that the treatment is likely to have some effect on crash frequency. However, there is unexplained systematic variation among the eight CMFs that is sufficiently large that they should not be combined into an overall average CMF for rural areas (or for rural areas).

Similarly, the standardized residuals are shown graphically at the top of the Main worksheet. Guidelines for using this figure to identify outliers are provided in the "Standardized Residual" section of the Introduction worksheet.

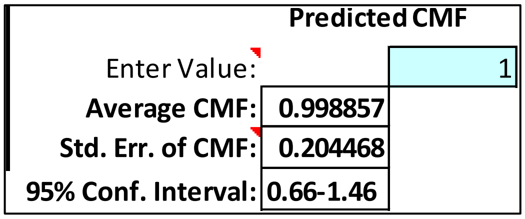
The overall average CMF for rural conditions is shown in Figure D12 as 0.659. The standard deviation of this CMF is 0.098 and the 95th percentile confidence interval is 0.49 to 0.87. Guidance for interpreting these statistics is provided in the section titled "Average CMF" of the Introduction worksheet.



	Predicted CMF	
Enter Value:		0
Average CMF:	**0.659363**	
Std. Err. of CMF:	**0.098127**	
95% Conf. Interval:	**0.49-0.87**	

Figure D12. Computed CMF for TWLTL installation in rural areas.

The CMF for urban conditions is computed by entering a “1” in cell N11. The value “1” is used because this value was defined as representing urban conditions using the indicator variable described in Step 3. The analyst then clicks on the Analyze Data button (as shown in Figure D6) to initiate the regression calculation process a second time. The overall average CMF for urban conditions is shown in Figure D13 as 0.999. The standard deviation of this CMF is 0.204 and the 95th percentile confidence interval is 0.66 to 1.46.



	Predicted CMF	
Enter Value:		1
Average CMF:	**0.998857**	
Std. Err. of CMF:	**0.204468**	
95% Conf. Interval:	**0.66-1.46**	

Figure D13. Computed CMF for TWLTL installation in urban areas.