NOVEC Data Analysis and Forecasting Cost Model Report

Grant Huang
Craig McClellan
Sajjad Taghiyeh
Junchao Zhuang

15 December 2015

Special Thanks to:

Robert Bisson, NOVEC

Bryan Barfield, NOVEC

Shayan Rashid, NOVEC

Dr. Karla Hoffman, GMU

Dr. Kuo-Chu Chang, GMU

Dr. Jie Xu, GMU

Table of Contents

Table of Figures

1 Background

Northern Virginia Electric Cooperative, as known as NOVEC, is responsible for the delivery of electric power to homes and businesses in a large portion of the Northern Virginia area. The company procures and distributes power to a multitude of commercial and residential customers.

NOVEC is a not-for-profit cooperative business headquartered in Manassas, Virginia. It is wholly owned by its member-owners. The company's service area includes the counties of Clarke, Fairfax, Fauquier, Loudoun, Prince William, and Stafford. It provides power for more than 155,000 residences and covers an area of 651 square miles. NOVEC maintains more than 6,880 miles of power lines [16].

Along with the maintenance of existing power lines and the servicing of everyday customer needs, NOVEC must plan, construct, and install new distribution services for the growing NOVA region. Figure 1 below illustrates the expected population change in Virginia counties from present day to 2040 [17].

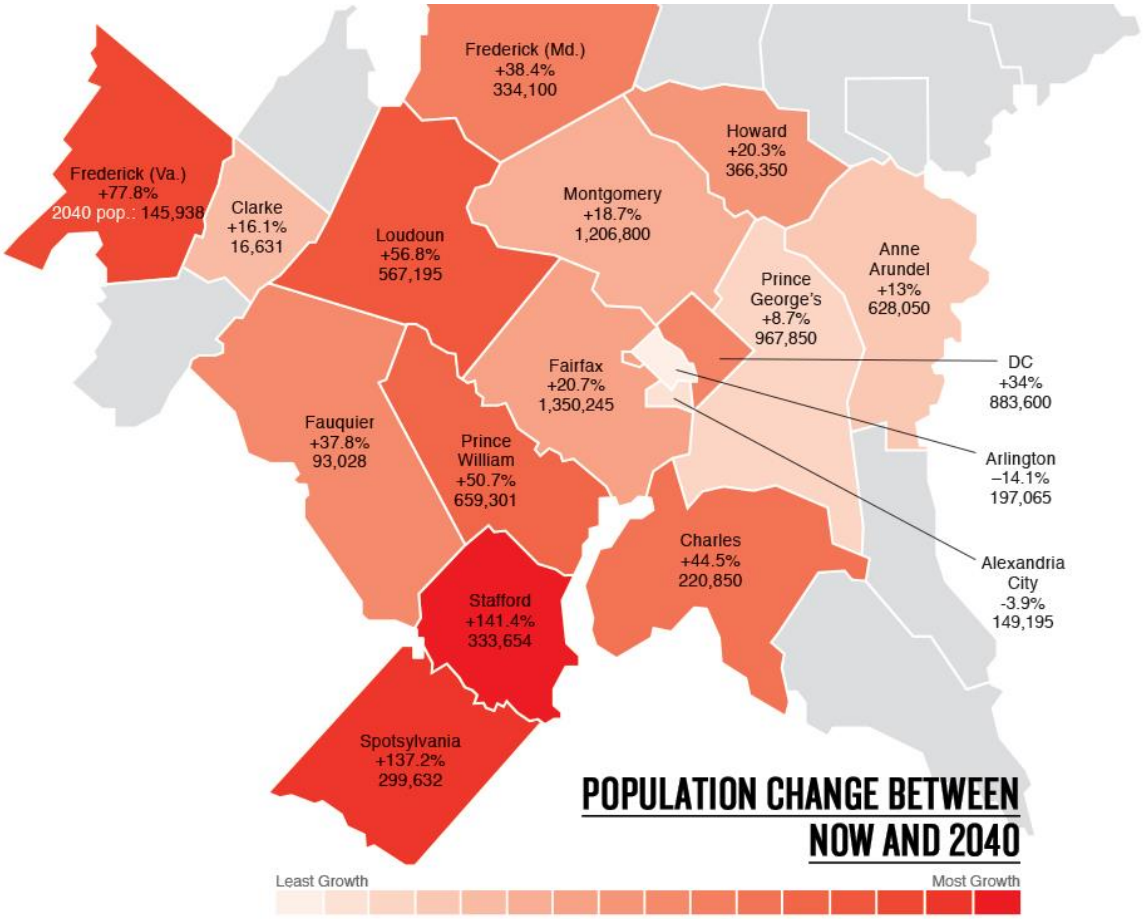

Figure 1. Population Change from Present to 2040

Projecting market demand accurately helps NOVEC to efficiently plan its construction activities and is likely to reduce the uncertainties associated with any procurement plans thereby increasing the likelihood that their costs will be reduced. Every year NOVEC uses D.C. region growth data to project a long term forecast of expected growth for their service area. However, forecasting construction costs associated with those new residential customers is very challenging.

2 Problem Statement

NOVEC has a rich historical record of the electrical utility construction projects that they have performed over the years. This collection of data has never been used to provide a short term prediction for residential construction costs.

At this time NOVEC does not have an analytic approach to forecast the costs they will incur for residential customers in the short term. The client asked our team to analyze their historical construction data and calibrate a model that will accept the forecasted number of homes (single-family homes, townhomes, etc.) and forecast the expected cost that will be incurred in a three year time frame.

This estimate will include the costs to connect the new residential customers to the grid. It should also include the costs for ancillary construction that is related to these residential customers. This model will be used to estimate the costs in a very near short term window that is assumed to be approximately three years.

The residential costs are assumed to be those associated with the model inputs. This is the number of new residential customers requiring connection to the grid for single family homes, condos, and townhomes.

The ancillary costs are those that are associated with Mainline, Infrastructure, “Other”, Barn, and Garage jobs or projects.

Mainline jobs are those that construct the large distribution lines in which other branches of the grid are fed. The number of these jobs are low but they represent a very substantial cost in the data.

Infrastructure jobs is a category that the team developed. This category is a consolidation of jobs related to Cable TV infrastructure, public lighting, traffic lights, and various other small projects. These jobs represent a very small amount of cost in the data set and small magnitude of jobs.

Other projects is a category description that is challenging to deal with. It represents a large amount of cost in the data set yet a relatively small magnitude in overall projects. This is a NOVEC named classification that does not provide a lot of insight into what these jobs actually represent.

Barn and garage projects represent jobs that connect new barns and garages to the electrical grid with their own meters. These can include free standing garages and barns that are remote relative to the regular electrical grid. These jobs represent a relatively small cost and small magnitude of projects.

3 Literature Review

This section provides a concise literature review on the methods used for this project.

3.1 Regression Analysis

Regression analysis is a technique which builds the statistical relationship between variables [1]. In order to relate the dependent variable to independent variable, it uses a mathematic model which can be defined as the following equation [2]:

$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon$$

Where Y is the dependent variable whose behavior depends on the values of $X_1, X_2, .., X_k$. The term ε describes the noise which is included in order to take into account the error which may be caused in predicting Y using X variables. There is a variety of applications of regression analysis in prediction and forecasting, which substantially is used in the field of machine learning. In regression analysis, it is investigated that among the independent variables, which of them are related to dependent variable and how they can be used in order to best explain the behavior of the dependent variable. Nevertheless, it may sometimes result in incorrect relationships, thus one should be cautious when using it [3]. There exist many techniques to perform regression analysis which in general are divided into two categories; Parametric and non-parametric. In parametric regression analysis, the regression equation is derived using a finite number of unknown parameters which are estimated from data. Linear regressions and ordinary least squares regression are among the parametric family of regression analysis. Non-parametric regression allows regression functions to rely on a specific set of functions, which may happen to be infinite dimensional. In practice, the form of the data generating process and its relationship to the regression approach has the most effect on the performance of the regression analysis. However, most of the times, the form of the data-generating process is unknown and hence, the regression analysis ends up making some assumptions regarding this process. Providing sufficient data, these assumptions can be validated. Even if these assumptions are moderately violated, the regression models used for prediction are still useful, but obviously not optimal. Nonetheless, the regression methods can provide false results [4][5].

In this project, a linear regression model was used as one of the methods to predict total cost. Then a cross-validation method was executed in order to examine the prediction power of our model. The following sections further explain these two methods.

3.2 Linear Regression Models

The linear regression is a parametric regression analysis used to model the relationship between one dependent variable (denoted as Y), and one or more independent variables (denoted as X). When there is only one independent variable, it is called simple linear regression. Additionally, when there is more than one independent variable, the regression analysis is called multiple linear regressions [6].

Linear regression utilizes the data to come up with the estimation for unknown model parameters and models the relationships using linear prediction functions [7]. Similar to other forms of regression analysis, linear regression uses the conditional probability of Y given X. In the cases which joint probability distribution of Y and X is investigated, it is called multivariate analysis.

When a dataset of n statistical units is provided, the relationship between the dependent and independent variable is assumed to be linear. Also, error variable ε is used to add noise to the linear relationship between the regressor and dependent variable.

3.3 Monte Carlo Simulation

The Monte Carlo Simulation is a term used to describe the techniques which utilizes statistical sampling in order to approximate the solutions to the quantitative problems. This technique is the most beneficial when other mathematical models are impossible to be used or it is hard to use them. The focus of Monte Carlo Simulation is mainly on three distinct classes of problems: optimization, numerical integration, and generating draws from a probability distribution [8].

Although there exist multiple Monte Carlo Simulation Methods, they mostly follow a specific pattern which is described as following four steps:

1) Defining the range of possible inputs
2) Using probability distribution to generate random inputs over the domain
3) Performing computation on the input
4) Utilizing all the results to generate required statistics

In this project, Monte Carlo Simulation is used as a tool to forecast the short term cost using historical data. Instead of drawing from a distribution, we sample from the historical data that is provided by the client. This procedure is called bootstrapping which is addressed in the following section.

3.4 Bootstrapping

In applied statistics, Monte Carlo and bootstrap method are frequently used as computer extensive methods. Bootstrap is considered to be a part of Monte Carlo family which is based on the observed data [9], [10]. Bootstrap was first introduced by Bradley Efron (1979), and since then, he has added quite a few specifications about this method and its generalizations. There exists a rich literature on bootstrap method which was mostly studied in the past two decades. These literatures demonstrate a wide range of applications for bootstrapping on real world problems.

In practical application concept, bootstrap refers to sampling with replacement from the actual data to generate bootstrap samples. The correct bootstrap sampling depends on the complexity of the data structure and it gets more complex as the complexity of the data structure increases. Bootstrap samples are considered to be a proxy for independent real samples from actual function. It worth mentioning that bootstrap can only work well for large sample sizes, and for small sample sizes the results may not be reliable.

3.5 Cross-Validation Methods

Simple cross-validation is the most popular and commonly incorporated cross-validation procedure. The approach begins with two subsamples randomly selected from the same sample. The first subsample is used as the calibration sample, and the second subsample is used as the validation sample. The regression model creates estimated regression coefficients according to the calibration sample. Applying these estimated regression coefficients in the validation sample data, a predicted value will be produced for the validation sample. In an ideal case, the validation sample should be collected separately from the data recorded in the calibration sample. Performing cross-validation of

a multiple regression model on a data set which is collected separately from the calibration data would alleviate the problem of capitalizing on chance occurrences that may have occurred during the data collection process. However, collecting new data is not always feasible, and it may lead to delays of the assessment of the multiple regression equation [11]. Therefore, typical cross-validation methods randomly split an available sample of data in half [12]. However, splitting the sample to yield both the calibration and validation samples are a serious drawback of the simple cross-validation procedure, especially for smaller samples, making the estimated regression coefficients not as precise as they would be if the entire sample of data were used when determining the regression model [13]. Since only half of the data is used to calculate the standard errors of the regression coefficients, the precision in the regression coefficients decreases and this precision decreases as sample size decreases. However, it is not possible to perform cross-validation if the entire sample is used to determine the regression model. [14]

The cross-validation technique which we use for our project is as follows:

Since we had the historical data for 10 years and we used 7 of them, we created 7 different regression models for purposes of cross-validation. In each of these regression and simulation models, we omit the data for the specific year (e.g. 2009) and use the rest of the dataset (e.g. 2008, and 2010 to 2014) to build our model. Then we use each model to predict the total cost for that specific year (e.g. 2009) and we compare the prediction values to the actual values.

Also in this project, we incorporated Autoregressive Integrated Moving Average (ARIMA) model to forecast the non-dwelling costs which we give a concise literature review about it here.

3.6 Autoregressive Integrated Moving Average (ARIMA)

In time series analysis, an autoregressive integrated moving average (ARIMA) model is considered as a generalization of an autoregressive moving average (ARIMA) model. These two time series estimation models are used either to understand the data more clearly or predict future series points (forecasting). They are implemented in the conditions that data show no evidence of stationarity. In these cases, in order to reduce the non-stationarity, an initial differencing step can be performed.

$ARIMA(p, d, q)$ refers to non-seasonal ARIMA models where parameters p, d, and q are non-negative integers, p is the order of the Autoregressive model, d is the degree of differencing, and q is the order of the Moving-average model. Seasonal ARIMA models are usually denoted $ARIMA(p, d, q)(P, D, Q)m$, where m is the number of periods in each season, and the uppercase P, D, Q denote the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model, respectively.[18][19]

3.6.1 ARIMA Definition

ARIMA(p' ,q) model is defined as follows:

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

Where X_t are the given time series data, L is the lag operator, α_i are the parameters of the autoregressive part of the model, θ_i are the parameters of the moving average part and ε_t are error terms. In general, it can be assumed that the error terms ε_t are independent, identically distributed (iid) variables sampled from a Gaussian distribution with mean equal to zero.

By assuming that the polynomial has a unitary root of multiplicity d, the first term in the left side of the previous equation can be rewritten as:

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) = \left(1 - \sum_{i=1}^{p'-d} \phi_i L^i\right) (1 - L)^d .$$

This polynomial factorization property can be expressed with $p = p' - d$ in an $ARIMA(p, d, q)$ process, which is shown in the following equation:

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

Hence, it can be considered as a particular case of an $ARIMA(p + d, q)$ process which has the autoregressive polynomial with *d* unit roots.

One can generalize the above equation as follows:

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) (1 - L)^d X_t = \delta + \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

Which gives the definition of an $ARIMA(p, d, q)$ process with $\boldsymbol{drift}$ $\delta/(1 - \Sigma\varphi i)$.

4 Data Analysis

4.1 Dataset Description

NOVEC kept the data for each construction project in a Work Management System database, which included the categories such as type, length, and cost of construction. The client provided the team with the construction data they had collected for the past 10 years. This data set consisted of over 300,000 individual pieces of data.

4.2 Challenges

4.2.1 Initial Observations of Dataset

Part of the scrubbing of the data, resulted in the client providing a second data set that removed some of the issues with the first data set. This second data set correct cost issues related to the labor and material costs associated with each project.

4.2.2 Difficulty in Distinguishing Between Commercial and Residential Data

Another issue with the data set was that it included some projects that appeared to be related to commercial constructions. Since the project is limited to residential construction, the team worked with client to determine which projects should be considered commercial and which should be considered residential. The initial attempt to reduce the dataset consisted of the team examining the WR_TYPE_DESC column. The WR_TYPE_DESC column consists of the following 35 entities:

- COMMERCIAL SERVICE
- COMMERCIAL SERVICE GRAPHIC
- CONDUIT SYSTEM
- DEMOLITION LETTER WORK REQUEST
- DIST LINE EQUIPMENT GRAPHIC
- DIST LINES GRAPHIC
- DISTRIBUTION LINE EQUIPMENT
- DISTRIBUTION LINES OVERHEAD
- DISTRIBUTION LINES UNDERGROUND
- MAIN LINE GRAPHIC
- MAIN LINE RESIDENTIAL OVERHEAD
- MAIN LINE RESIDENTIAL UNDRGRND
- MAINLINE CABLE PULL
- OH COMM SRV SMALL
- OH RESIDENTIAL SERVICE
- POWER SUPPLY GRAPHIC
- POWER SUPPLY
- RESID SERVICE - CABLE PULL
- ROAD CROSSINGS
- SERVICE GRAPHIC
- SERVICE SUBD GRAPHIC (NO DRAW)
- SHORT RANGE PLAN WORK GRAPHIC

- SHORT RANGE PLAN WORK
- STREET LIGHTING (COUNTY)
- STREET LIGHTING (OTHER TYPES)
- STREET LIGHTING COUNTY GRAPHIC
- STREET LIGHTING OTHER GRAPHIC
- TEMPORARY SERVICE
- TEMPORARY SERVICE GRAPHIC
- UG COMM SRV LARGE
- UG COMM SRV SMALL
- UG RESID SRV NO ENGINEERING
- UG RESID SRV W/OH SPAN
- UG RESIDENTIAL SERVICE
- VDOT ROAD IMPROVEMENT

After conferring with the client, it was determined that the determination should be based on the “CLASSIFICATION_DESC” column rather than the “WR-TYPE_DESC”. The CLASSIFICATION_DESC column contains the following 28 entities:

- APARTMENT
- BARN
- BETHLEHEM SUBSTATION
- BROAD RUN SUBSTATION
- CABLE TV
- CONDO
- COUNTY LIGHT
- FAST FOOD RESTAURANT
- FOOD STORE
- GARAGE
- HOA LIGHT
- HOTEL
- MAINLINE
- MOBILE HOME
- OFFICE
- OTHER
- PRIVATE HOMEOWNER LIGHT
- RESTAURANT (OTHER)
- SALES TRAILER
- SCHOOL
- SGL FAMILY HOME
- STRIP MALL
- TELEPHONE
- TOWNHOUSE
- TRAFFIC SIGNAL
- VA DEPT OF TRANSP

- WAREHOUSE
- WORK PLAN

The team then assessed the CLASSIFICATION_DESC column and again developed a list along with assumptions to differentiate commercial and residential items. The challenge of clearly knowing if an entity belongs in commercial or residential still exists. There are many entities where the team believed to belong in both groups. This list was sent to the client for review and feedback was provided to assist us in reducing the dataset.

4.3 Data Normalization

The historical data set that NOVEC provided represented costs from 2005 to 2015. We agreed that an apples to apples comparison of costs required a net present value normalization. The team initially attempted to do this conversion by using the electrical price inflation over the span of years that concerned our data set. Upon consultation with NOVEC though we were alerted that this was significantly different than electrical utility construction during those years.

To aid with normalization, NOVEC provided the Handy Whitman Index of Construction Costs [15]. This reference contains the costs for materials and labor for elements of the construction industry. It is broken out by many different types of construction. It also collects this information by the different regions of the United States. This provided a very specialized collection of data that was utilized to create inflation normalization factors.

We focused on the electrical utility construction data of the South Atlantic Region, obviously, due to this is being NOVEC's region of business. The subset of data that we selected was a varied "basket of goods" that encapsulated the exact type of construction that we are concerned with.

Handy Whitman Electric Utility Construction Cost Basket of Goods
Poles, Towers, & Fixtures
Overhead Conductors & Devices
Underground Conduit
Underground Conductors & Devices
Line Transformers
Services - Overhead
Services - Underground
Meters Installed
Street Lighting-Overhead
Mast Arms & Luminaires Installed
Street Lighting-Underground

Figure 2. Handy Whitman Electric Utility Construction Cost Basket of Goods

The data used from the Index mirrored many of the line items that were observed in the unformatted data set. This provided confidence that inflation factors based on this data would be a very good approximation of inflation that the company actually experienced over the time frame of the dataset.

The cost for the complete basket of goods were summed per year. Then the delta was calculated from the previous year and divided by the sum of costs from the previous year. This resulted in the percent change per year of the cost.

Electricity Data			Consumer Inflation		
Source	Year	Year to Year	Source	Year	Year to Year
South Atlantic Region Distribution Inflation	2005	7.81%	Consumer Price Index	2005	3.40%
	2006	8.85%		2006	3.20%
	2007	13.92%		2007	2.80%
	2008	5.05%		2008	3.80%
	2009	10.06%		2009	-0.40%
	2010	-0.89%		2010	1.60%
	2011	1.32%		2011	3.20%
	2012	4.95%		2012	2.10%
	2013	2.24%		2013	1.50%
	2014	-0.43%		2014	1.60%
	2015	1.13%		2015	0.20%
Total Average:		4.91%	Total Average:		2.09%

Figure 3. 2005 - 2015 Electrical Utility Construction Inflation Compared to Consumer Inflation [20]

The table above shows the calculated inflation as compared to the Consumer Price Index during the same time frame. The inflation during the early part of the time frame is very high compared to regular inflation. The team had many theories as to why. It was simply possibly that a boom in construction led to this high inflation. We also suspected that the cost for materials could be influenced by the need for reconstruction in areas affected by hurricane Katrina. No matter the cause, it was obvious normalizing the data would be vitally important.

Next inflation factors were calculated that would inflate all the dollars to base year (BY) 2015 dollars. We started by deflating 2014 dollars to 2015 dollars and worked backward. The factors that were calculated are below in Figure 4. To calculate base year 15 dollars for dollars in a previous year you simply divide by the factor that represents that year's inflation factor.

BY Factors	Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
	BY Factor	0.6429	0.6998	0.7972	0.8374	0.9216	0.9135	0.9255	0.9714	0.9931	0.9888	1.0000
	How to use: To Convert 2015 dollars to 20XX dollars multiply by factor, to convert 20XX dollars to 2015 dollars divide by factor											

Figure 4. Calculated Electrical Utility Construction Inflation Factors

With the factors in hand Microsoft Excel was leveraged to inflate all of the costs in the data set to base year 2015 dollars. The analysis could now be done on a normalized data set.

5 Cost Driver Analysis and Data Formatting

The initial data set was an amalgam of both commercial and residential construction jobs. We worked with the client to identify those line items, and jobs that should be removed due to being more commercially related. These included items such as restaurants, substations, and other projects that would corrupt the analysis we were trying to perform. The data set required proper formatting before any analysis could be done for regression modeling. The original data, once cleaned of commercial data, consisted of approximately 300,000 line items that represented cost elements of all the residential related jobs from 2005 to 2015.

Our initial analysis looked at the data set as a whole as we tried to ascertain which cost line items of each job were the most important.

In order to do this we first wanted to identify the real cost drivers of the projects. The data set was broken down into many varying levels of granularity of cost. Some of these were a challenge for us to understand. Admittedly our knowledge of the Electricity Utility Construction business is not very vast. We do have some knowledge of commercial construction and that along with regular consultations with NOVEC did clear up many of the questions we had.

We identified that the columns labeled Category Code identified each line Item as a type of construction. We also identified that within those groups were subgroupings labeled in the Catalog ID that further broke down these groups.

The Category Code was investigated first. Figure 5 shows the total breakout percentages for the complete data set of the Category Code Items. It is observed that the cost of Conductor, Trenching, Transformers, Labor, and Conduit are the major cost drivers for the data set.

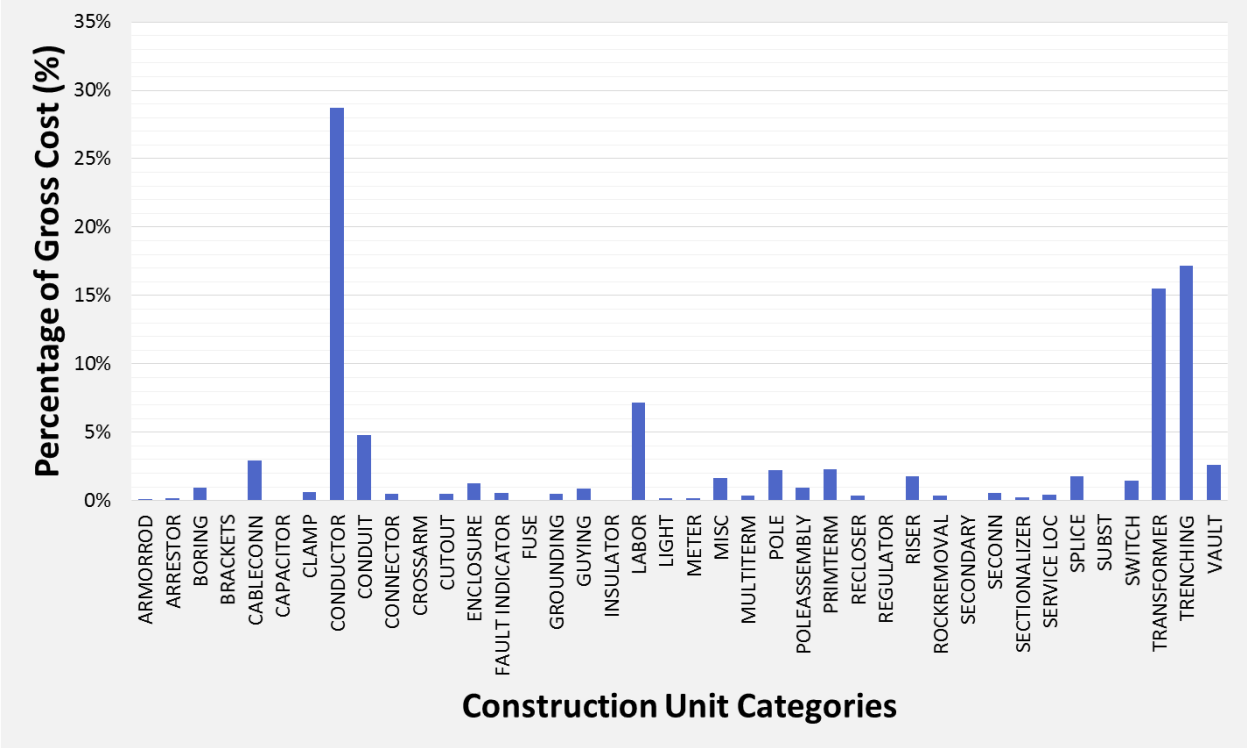

Figure 5. Percentage of Gross Cost Breakdown by Construction Unit Categories

This provided the first initial clues as to what technical parameters the team would need to investigate as predictors.

We knew initially that it would be desirable to include the number of meters that each job connected as a predictor of cost. Speaking with our client we knew that the input to the model would be a prediction of new residential customers that are going to be provided an initial meter installation or service. The number of meters allocated to each job would be a representation of the number of new home types that were serviced in each job. We hoped that a good relationship between the magnitude of meters in each job and its cost would be a "simpler" model to calibrate.

The Catalog ID represents a further breakdown in granularity to the Category Code data breakouts.

CU_CATEGORY_CODE	CU_CATALOG_ID	%
CONDUCTOR	CABLE	28%
TRANSFORMER	1PHASE	15%
TRENCHING	MACHINE	14%
LABOR	LABOR	7%
CONDUIT	PVC	4%

Figure 6. Category Code Breakdown by Catalog ID

Figure 6 above shows that the Conductor Category code is made up entirely of the cable Catalog ID. The majority of the Transformer cost is made up of 1 Phase type transformers. Trenching has several lower level cost elements but the largest contributing cost driver is trenching in which a machine is required. It is observed that the top three cost drivers represent over 50% of the total cost of the entire data set.

This analysis tailored the way ahead to format the data set so that it could analyzed. It was decided to construct the data set by rolling up the costs for labor, material, and overhead for each individual cost item. We also would sum up values for meters, conductor, transformers, machine trenching, labor, and finally conduit.

Machine trenching is the major source of cost for the trenching Catalog Code. Other line items include hand trenching, bedding, and others. These line items represent a much lower magnitude of the costs the makes up trenching cost items. Many of the trenching items have different units. We had to assume that adding up all of the separate type of items that had units of measure (UOM) in feet could perhaps be double counting the same stretches of trench. In this way machine trenching, which represented the majority of the cost, would be a proxy for the other aspects of trenching related to the same trench.

Our completed format of vectors contained a rollup of each job that contained the Work Number, Date Completed, Job Type (Single Family home, Mainline, Infrastructure, etc.), Overhead or Underground Job Classification, Sum of Meters, Labor Cost, Material Cost, Overhead Cost, Contribution. It has columns that further label the Feet of Conductor, Feet of Machine Trenching, Labor, and Number of Transformers. The baseline costs are calculated based on the sum of the labor, material, and overhead costs. (Appendix C: Refined Data Vector Examples)

6 Data Selection

After normalizing the data set to base year 2015 dollars, the team attempted to identify any patterns or trends to show the relationship between residential construction cost and time.

The first observation, as shown in the pie chart below, demonstrated the percentage of gross cost by job classification. Mainline projects, "Other" projects, and "Single family home" projects were the top three classifications, totaling 83% of the total gross cost. For Mainline projects, the number of projects were relatively small at 7% of the total job counts, but each of them had relatively high cost due to the construction scope. It was unclear what specific activities were included in the "Other" classification. Perhaps they are made up of both residential and commercial activities. We know that they contain a relatively high amount of high cost projects. Single family home projects, due to their large magnitude of the job counts, 79% in total, has the third largest in total gross cost.

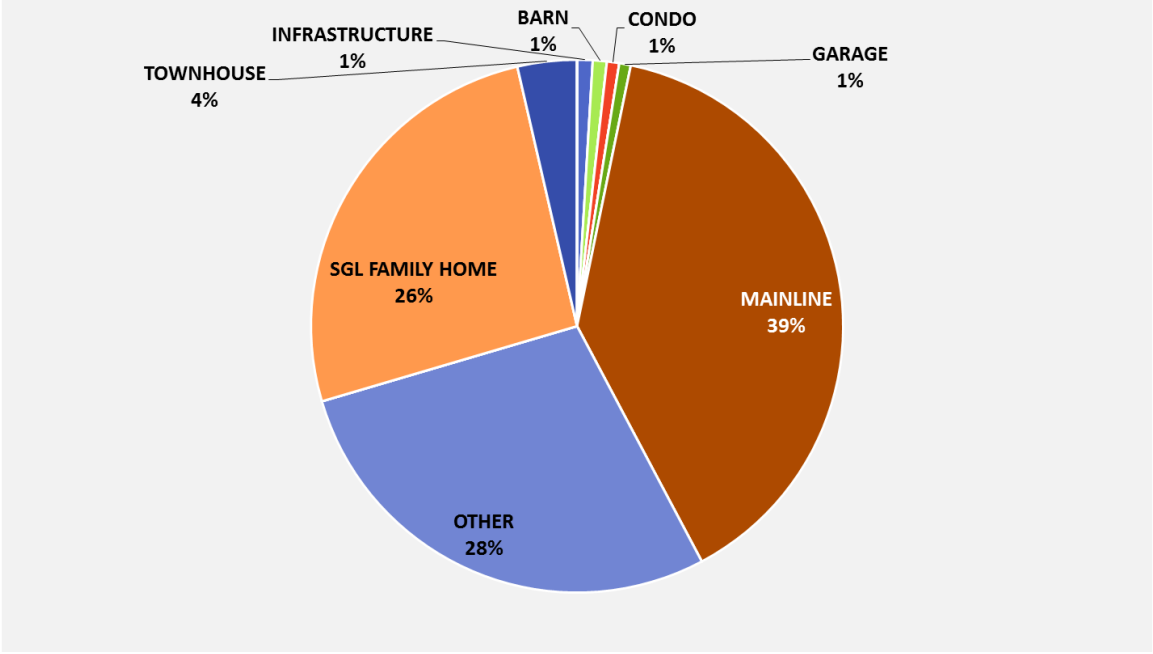

Figure 7. Percentage of Gross Cost by Job Classification

The second observation was that during the analysis of the annual gross cost, the team noticed that the gross costs from 2005 to 2007 was significantly higher than other years. This behavior was quite different when compared to the gross costs from 2008 to 2014. From 2008 onward, the decline disappeared and the gross costs for the following years gradually increased with a consistent rate.

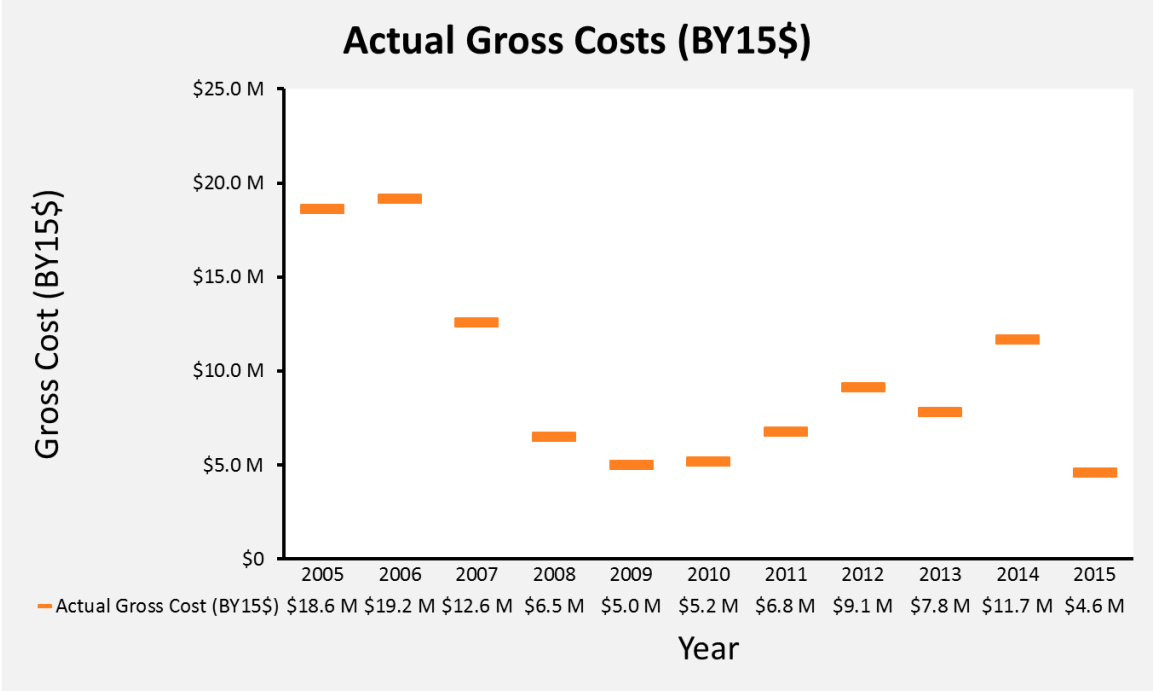

Figure 8. Actual Gross Costs (BY15$)

In general, with a higher total number of construction projects, as shown in Figure 9, it is reasonable to have a high gross cost in the early years. However, the team observed that a relatively large portion of the expensive projects occurring in the early years. Looking at the top 50 projects in 2005 to 2007 with respect to cost for each job classification, the top three classifications in percentage of total gross cost are single-family homes, Mainline projects, and "Other" projects. They are 52%, 66%, and 72% respectively as demonstrated in Figure 10. This is demonstrated in Figure 9 below. This phenomenon matches the performance of the housing market and economy during the period of time.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Number of House	13059	9607	6389	3993	3806	4189	4532	5441	5637	5974
Barn Project	31	33	16	10	9	14	17	16	13	11
Garage Project	36	22	31	17	19	10	13	14	14	9
Mainline Project	42	99	75	68	54	80	102	118	111	140
Infrastructure Project	55	59	71	33	23	13	25	10	17	34
Other Project	244	234	186	170	175	118	122	141	139	147

Figure 9. Historical Number of Projects and Number of Homes

Top 50 Cost Project in	2005~2007	2008~2014
S.F.H.	52%	48%
T.H.	20%	80%
Condo	28%	72%
Infrastructure	28%	72%
Barn	24%	76%
Garage	32%	68%
Mainline	66%	34%
Other	72%	28%

Figure 10. Distribution of Top Costing Projects: 2005 – 2007 vs 2008 – 2014

The analysis performed at the level of specific types of projects showed that the average costs for mainline and other projects were relative high in the early years with rapidly decreasing trends. Meanwhile, other types of construction did not these relatively large variances. Since the mainline projects and the other projects were the two largest portion of the total gross cost, 39% and 28% respectively, they greatly influence the annual cost. We had to assume that in order to conduct the forecasting, the cost trend from 2005 to 2007 was abnormal and not suitable to predict cost in short term. All the data points in that three-year period were removed from further analysis and forecast.

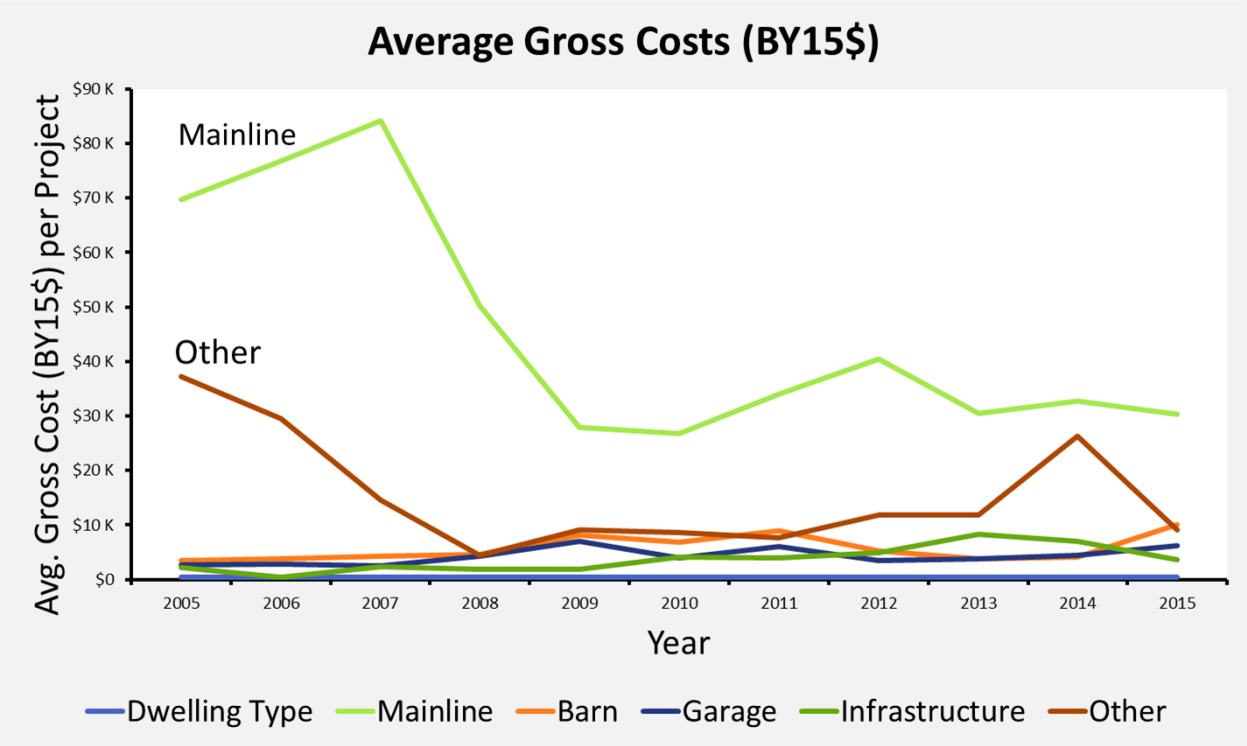

Figure 11. Average Gross Costs (BY15$)

7 Regression Model

For the regression based model, originally the entire dataset was considered which consisted of data ranging from the years of 2005 to 2015. The regression equations, models, and validation results that were developed corresponding to this data set time frame are located in section 16.2.

However, as mentioned in the Data Selection section, the dataset ranging from the years of 2008 to 2014 was deemed more suitable. As a result, the following material pertains to the arrived regression based model solution. Supplementary regression equations, models, and validation results generated from the 2008 - 2014 dataset are located in section 16.3.

7.1 Regression Model Approach and Algorithm

Leveraging R programming language to facilitate the data analysis process, the many faces of regression were explored in conjunction with various predictor variables. These types of regression included simple linear, polynomial, multiple linear, and logarithmic transformations.

Utilizing the reduced data set that pertained to residential costs, the data was further separated by jobs that pertained to one of the home types (single-family home, townhome, and condo) and the remaining job classifications (Barn, Garage, Mainline, Infrastructure, and Other).

Given that the input to the model will be the projected number of residential customers, the initial rounds of regressions relied on using the number of homes as the predictor variable. This would have been ideal and the most practical solution for NOVEC. However, the analysis results indicated that there was low correlation between gross cost and the number of homes. This required the team to take a deep dive and investigate top cost drivers for jobs. The pie chart presented below in Figure 12 illustrates the top job cost drivers.

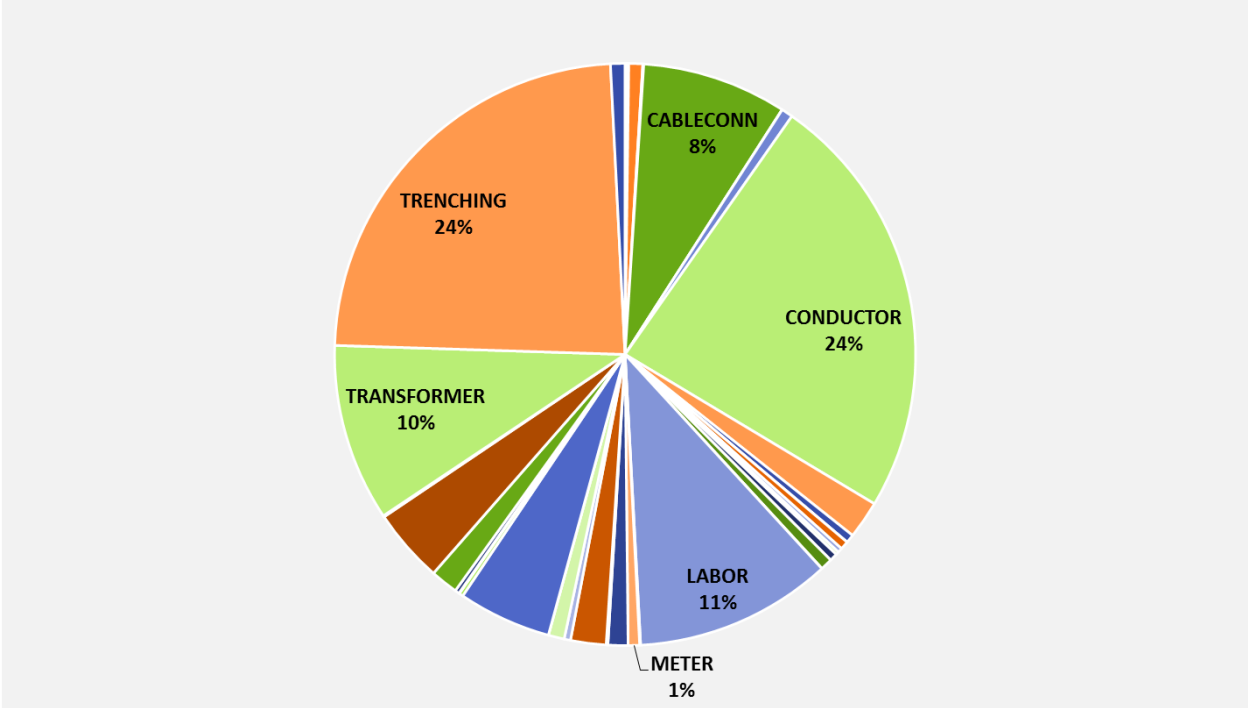

Figure 12. Job Cost Drivers

After analyzing the sequential rounds of regressions that were based on the top cost drivers, the length of conductor cable parameter was selected as the predictor variable for a number of reasons. It is a major cost driver and has a solid linear relationship to job gross cost because the distance a home is away from a mainline is correlated to cost. Additionally, this technical parameter was recorded the most consistently with respect to the labeling of units and number of observations. Lastly, this is a practical solution. The best regression based model is centered on the length of conductor cable as the predictor variable. This model was evaluated through an initial validation stage and further evaluated by cross validation.

As shown in Figure 13 below, the regression model algorithm takes the projected number of residential customers for some future year provided by NOVEC and outputs a point value forecast along with a predicted range in base year 2015 dollars.

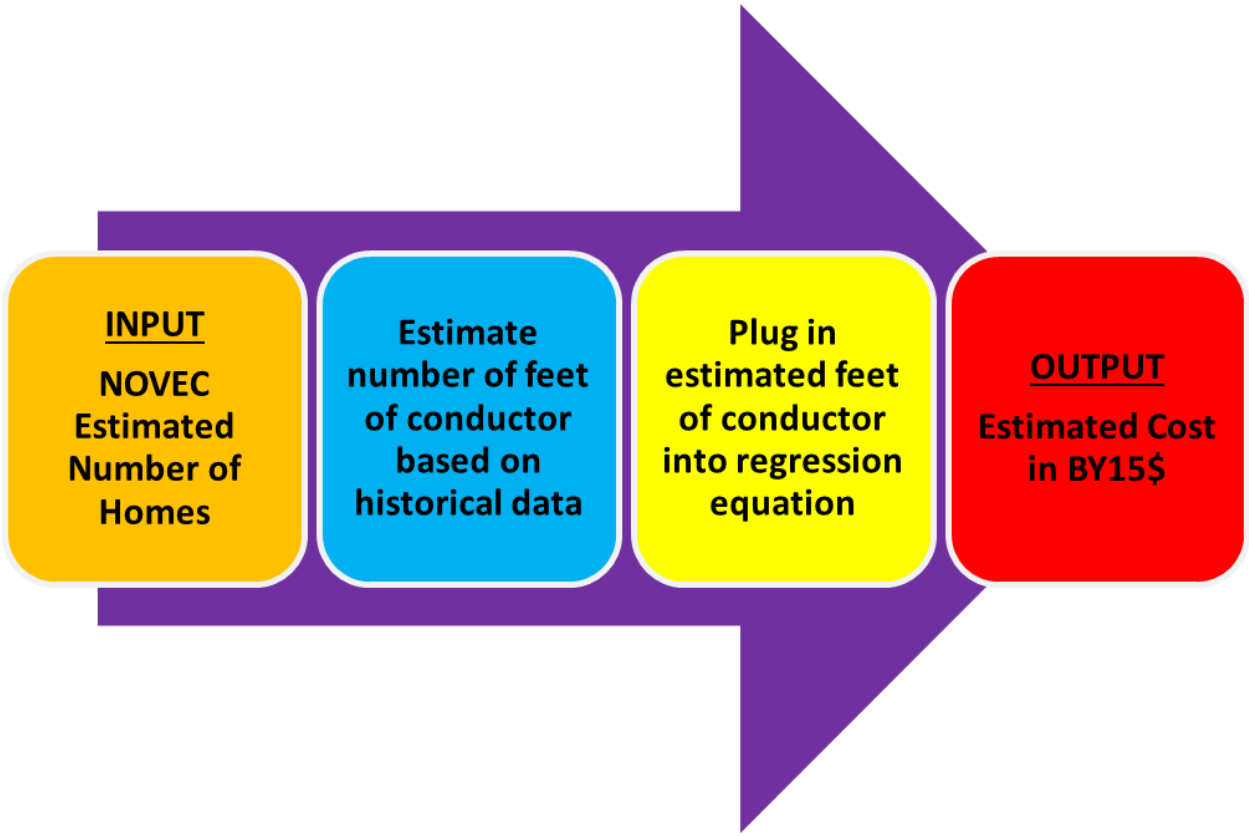

Figure 13. Regression Model Algorithm

The process in between is a two-fold. Since the number of meters is a proxy for the number of homes, the job distributions that correlate to a unique number of homes were determined based on historical data. Furthermore, the job distributions that correlate to a unique number of length of conductor cable are determined as well. Employing both of these distributions, NOVEC'S estimated number of homes is transformed to total number of jobs and then converted to the number of jobs that are associated with each distinct value of length of conductor.

The second phase is where the costs associated with each unique value of length of conductor cable are computed with the regression equation. The sum product of the number of jobs and the cost that correspond to each exclusive value of length of conductor cable from the regression equation results in the point value prediction of cost for some future year in base year 2015 dollars.

7.2 Cost Burdening

The regression model is similar to the stochastic simulation model in terms that the input will be some value of predicted homes. This single input value represents a sum of single-family homes, condos, and townhomes altogether. The main distinction between the two different model approaches is with respect to how the cost of those jobs that do not belong to one of the home types are handled. The regression approach treats the costs for Mainline, Other, Barns, Garages, and Infrastructure as a burden that is incurred in order to bring new residential customers into the grid. The assertion is that these extra costs are the "costs of doing business". All of these elements are necessary whenever a new home or development is constructed. Therefore, these costs need to be accounted for and somehow allocated to the cost of the home type jobs, referred to as the baseline gross cost. The summation of baseline gross costs and burden costs makes up the response variable in the regression model.

Many challenges were faced when exploring techniques to properly allocate the burden costs to baseline gross costs. The first challenge relates to not knowing how those jobs within the burden cost bucket are correlated with homes in terms of completion dates. As a result, the assumption was made to allocate the total sum of the burden costs per year to all of the home type jobs that occur in that same year. It is both recognized and understood that this may not be totally correct since Mainline and Infrastructure jobs may not be constructed within the same completion year with the homes that they support. The assumption is that this will even out for every year. Figure 14 below presents the percentage of both baseline gross costs and burden costs for each of the 7 years. Baseline gross costs range from 25% to 35% and burden costs range from 65% to 75%. Note that burden costs are quite significant and are two times as much as the baseline gross cost on average.

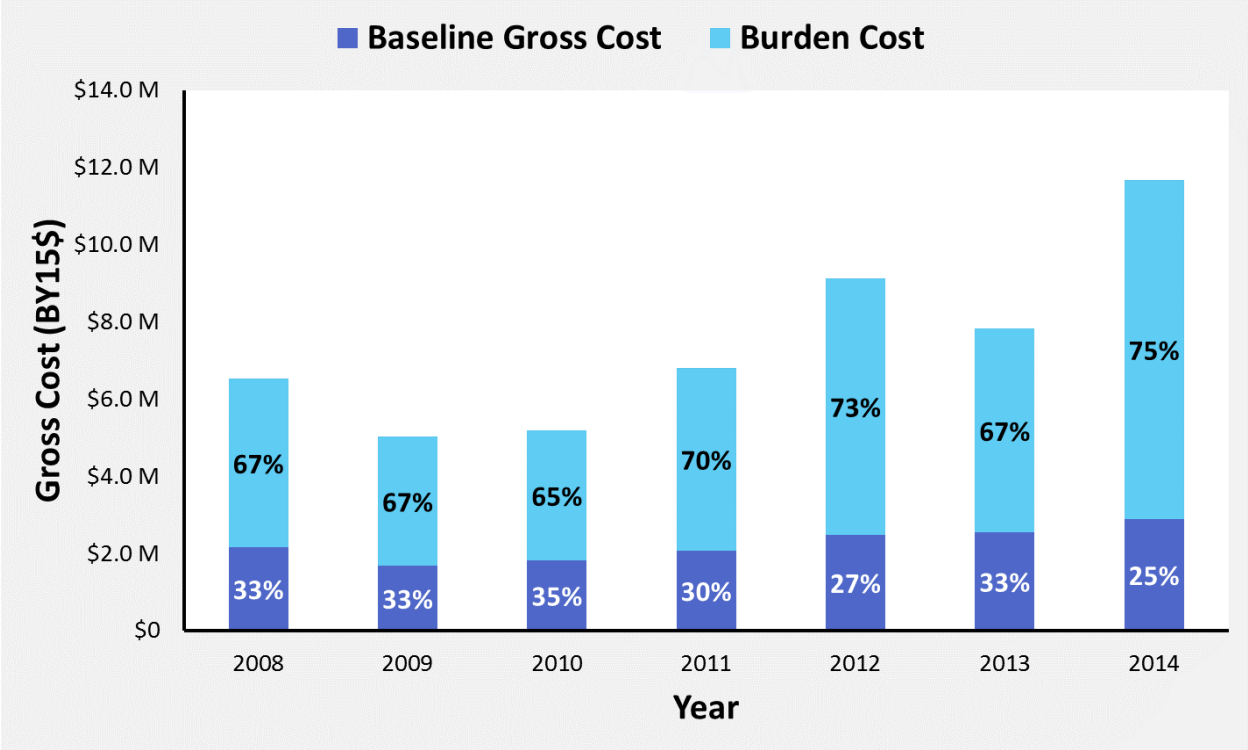

Figure 14. Baseline Job Cost and Burden Costs per Year from 2008-2014

The next challenge was to determine how to exactly allocate the total burden cost per year to the home type jobs that were performed that year. Three different avenues were explored to achieve this.

The first allocation method evenly spreads the burden costs to every job, denoted as equally distributed. The total burden cost for each year is divided by the number of jobs that were performed for that year and the resultant cost is added to the baseline gross cost of each home type job.

The next method employed was based on the number of meters that corresponded to each home type job. The notion was that the more meters a job required to be installed, more of the incurred burden should be allocated. Therefore, jobs with a high percentage of meters for the year would be given more burden costs. The total number of meters per year was calculated and each job received a percentage of that year's burden based on the job's meters divided by the total quantity of meters for that year.

The final selected allocation method was based on the weight of a home type job's baseline gross cost. The concept behind this method is revolved around being as fair as possible. In the original data set, several jobs included line items that described a contribution amount that was a reimbursable from the customers for the work completed. The baseline gross cost for each job is considered to be the sum of costs associated with labor, material, and overhead. We did not adjust the base gross cost to account for the contribution from the customer that would have reduced the gross base cost to a net base cost. The percentage of allocation for each job was computed by dividing the job's baseline gross cost by the total gross baseline gross costs of each home type job in

the job's completion year. This was done to keep the proportions of each job's baseline gross cost intact relative to the other job's baseline gross costs in each the year. The resultant allocation was added to every home type's job base gross cost.

Each type of allocation resulted in 3 different gross costs and regressions were ran against these three independent variables. This information is documented in Appendix D: Other Regressions.

The selected regression equation is located below in Figure 15 to obtain costs from the predictor variable of length of conductor cable.

$$\boldsymbol{Gross\ Cost\ (BY15\$) = (6.48475 \times Length\ of\ Conductor\ Cable) + 918.12091}$$

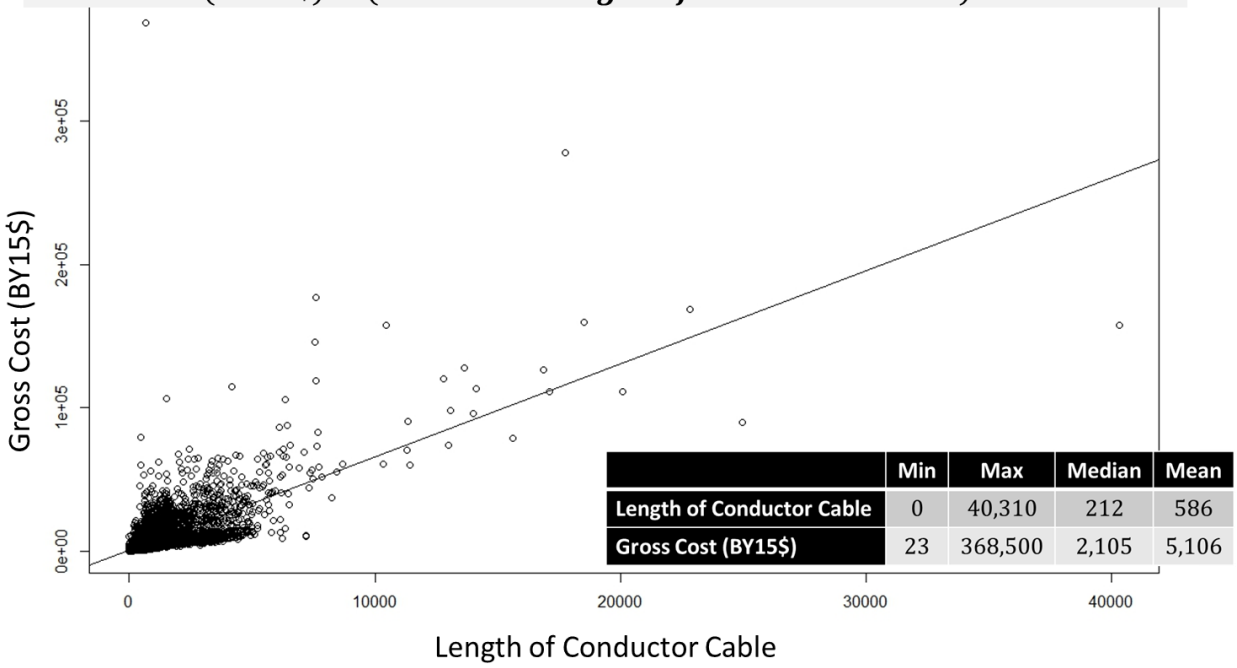

Figure 15. Linear Regression Equation for Regression Model

7.3 Correlation Analysis

In attempt to format the data in such a way that analysis could be performed efficiently, the team found many clues that helped pare down the data to a set of parameters that should be investigated.

Once the data was properly formatted, both Microsoft Excel and R were leveraged to perform a correlation analysis on the data. We wanted to see how the technical parameters of the jobs correlated to the different burdened costs.

Together with cost driver analysis we identified what we suspected to be the most important technical parameters. Figure 16 below shows that Meters, Feet of Conductor Cable, Machine Trenching, Labor, the Number of transformers, and Cable Connectors are correlated at various degrees with the different burden types

TCE	Gross Cost with Burden (Equally Distributed)
TCM	Gross Cost with Burden (Meters)
TCG	Gross Cost with Burden (Gross Cost)
M	Meters
CC	Conductor Cable
TRM	Trenching Machine
L	Labor
TF	Transformers
CA	Cable Conn

	TCE	TCM	TCG	M	CC	TRM	L	TF	CA
TCE	1.0000000	0.6736882	0.9577598	0.196223676	0.7522166	0.677560316	0.59241357	0.60594655	0.23816360
TCM	0.6736882	1.0000000	0.6607166	0.838222959	0.6907461	0.364484928	0.38441557	0.35392472	0.66923938
TCG	0.9577598	0.6607166	1.0000000	0.193170177	0.7731258	0.701953731	0.61529586	0.62791869	0.24651885
M	0.1962237	0.8382230	0.1931702	1.000000000	0.3900455	-0.001048784	0.08949195	0.03774925	0.75096145
CC	0.7522166	0.6907461	0.7731258	0.390045500	1.0000000	0.671724717	0.36885041	0.44382086	0.48496310
TRM	0.6775603	0.3644849	0.7019537	-0.001048784	0.6717247	1.000000000	0.35882945	0.47245672	0.05554290
L	0.5924136	0.3844156	0.6152959	0.089491950	0.3688504	0.358829451	1.00000000	0.46140926	0.09484274
TF	0.6059465	0.3539247	0.6279187	0.037749252	0.4438209	0.472456717	0.46140926	1.00000000	0.10430612
CA	0.2381636	0.6692394	0.2465189	0.750961455	0.4849631	0.055542896	0.09484274	0.10430612	1.00000000

Figure 16. Correlation Matrix

From this, regressions could be constructed over all of the permutations of dependent variables and independent variables. The correlation analysis was only used to select the variables in which further analysis should be performed.

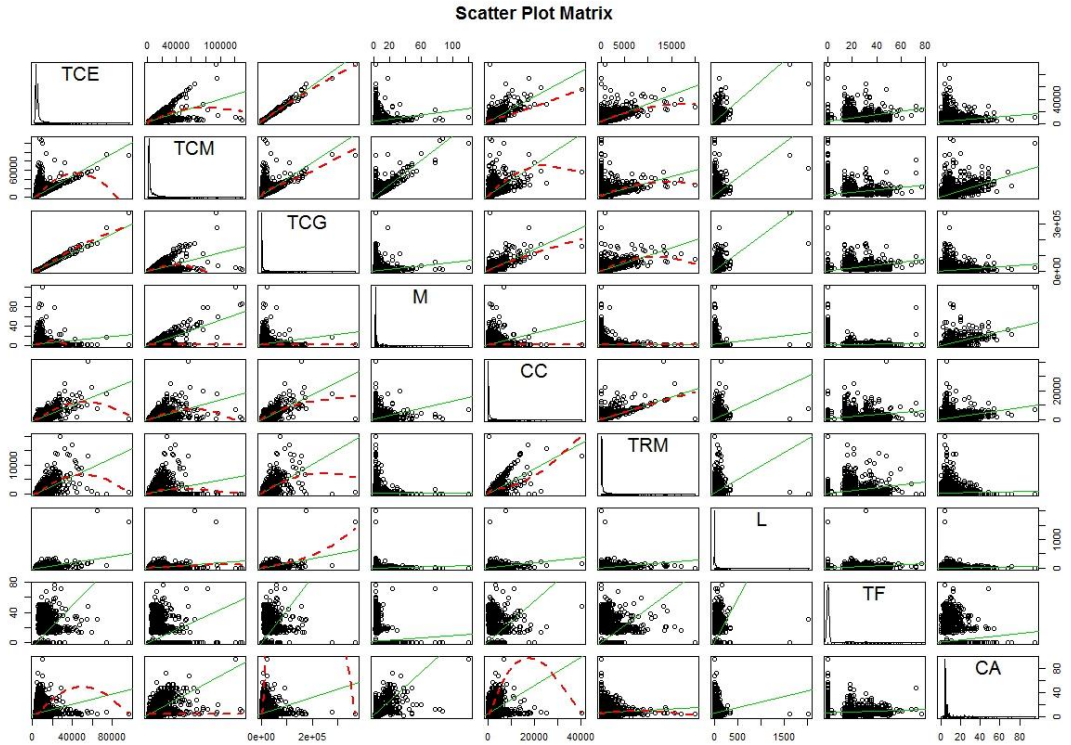

Figure 17. Correlation Matrix Plot

7.4 Validation

After the development of the model, as an initial validation the entire training dataset was used to forecast costs for the years of 2008 to 2014. The comparison of the predicted point value forecasts against the actual gross costs are illustrated below in Figure 18. The bars represent the residuals where the positive values indicate cases of overestimation and the negative values represent instances of underestimation.

7.4.1 Initial Validation

After the development of the model, as an initial validation the entire training data set was used to forecast costs for the years of 2008 to 2014. The comparison of the predicted point value forecasts against the actual gross costs are illustrated below in Figure 10. The bars represent the residuals where the positive values indicate cases of overestimation and the negative values represent instances of underestimation.

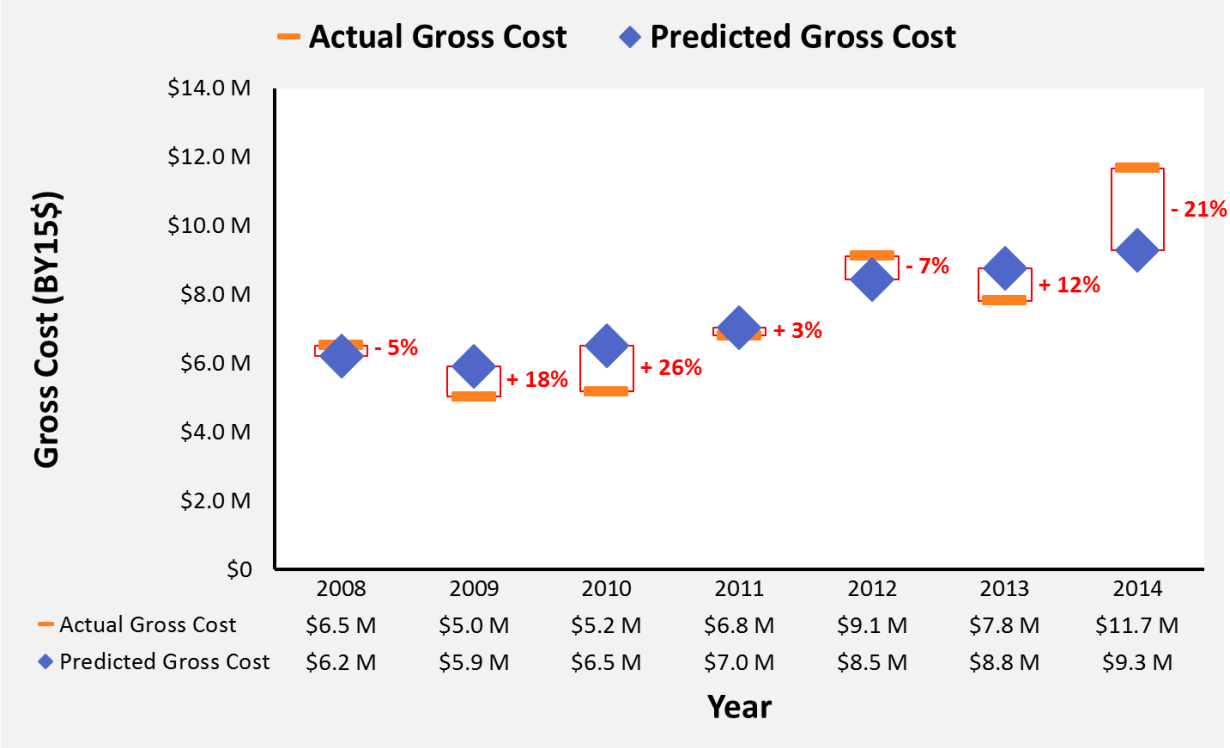

Figure 18. Initial Validation Graph

This initial validation results indicate that the worst case scenario in the positive direction occurs in 2010 at 26% and the worst scenario in the negative direction occurs in 2014 at 21%.

7.4.2 Cross Validation

As shown in Figure 18, the results from the initial validation were on the entire training data set which the learner had already seen. The concern with just relying on this technique to evaluate a model is that it does demonstrate how well the learner will perform when making new forecasts on data that it is not already seen. Taking it a step further, the model evaluation technique of cross validation was used to measure how accurately the model would perform in reality. The theory behind this method is to separate the data into training and testing sets. In

our case, specifically the holdout method was implemented where a single year of data was partitioned as the testing set and the remaining 6 years of data as the training set. For example, in the first iteration 2008 was used as the testing set and the data from 2009 to 2014 was used to train the model. This process was replicated so that each year had an opportunity to act as the testing set. A total of 7 iterations were performed, which resulted in newly generated regression equations each time. The results from cross validation are shown below in Figure 19.

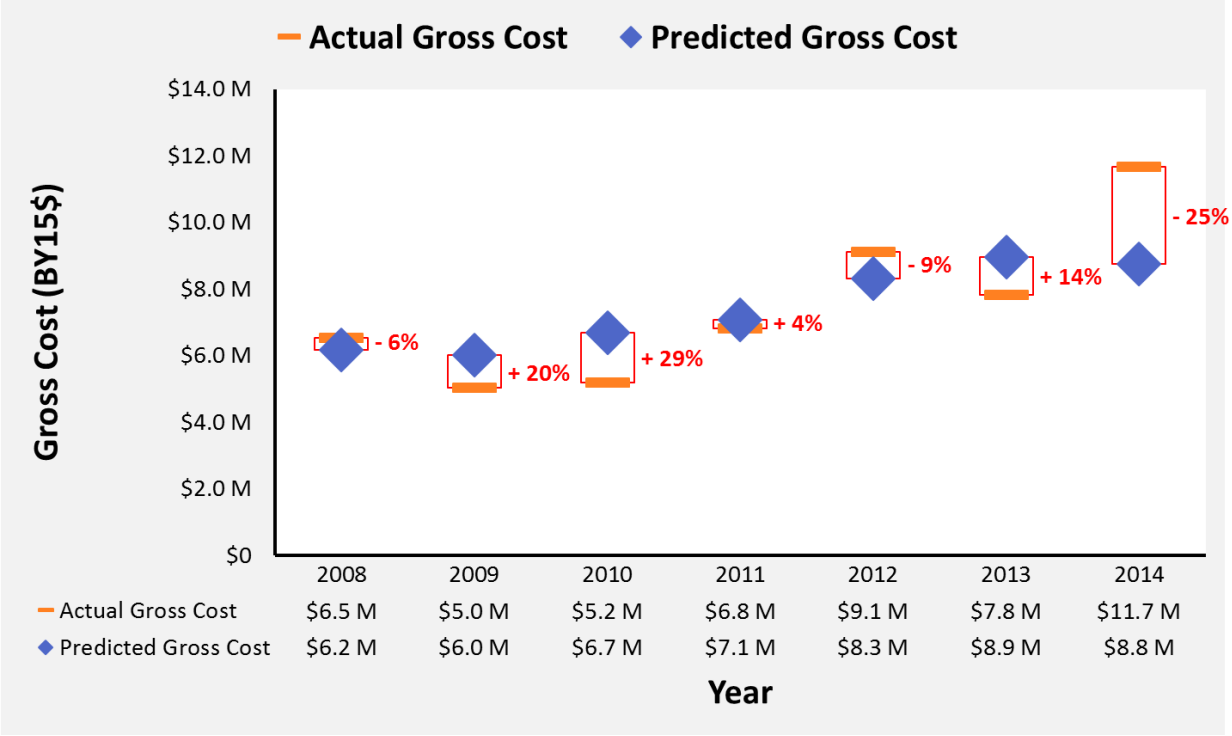

Figure 19. Cross Validation Graph

In 2010, the worst case scenario is observed in the positive direction at 29% and the worst case scenario in the negative direction is 25% in 2014.

7.4.3 Comparison of Residuals: Initial Validation vs Cross Validation

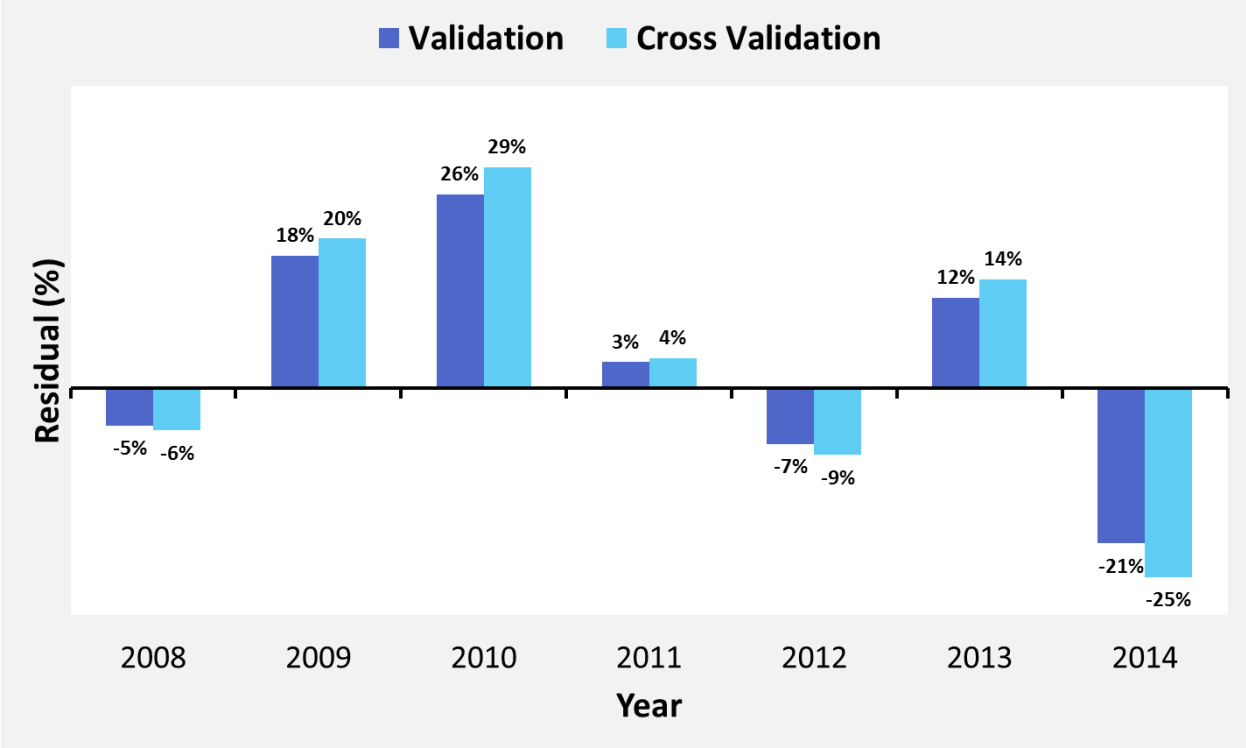

Figure 20. Comparison of Residuals Graph

As the residuals are compared from the initial validation and cross validation across the 7 years, the results are very similar. As expected, the residuals from the cross validation are worse, but the delta is minute. The differences in residuals range from 1% to 4%. By taking the worst case scenarios from the cross validation, a range estimate can be produced alongside the point value forecast.

7.5 Regression Conclusion

Referring back to the breakdown of costs by classification in Figure 7, it is vital to recognize that there are substantial costs in the burden cost bucket that are not well understood with respect to how they relate to homes. For instance, Mainline jobs account for 39% and “Other” jobs are 28% of overall gross costs. The combination of these two job classifications within the burden cost comprises over half of overall gross costs at 67%.

In an attempt to create a regression based model, this is the best model that could be achieved with the provided data set. Figure 19 illustrates the regression predicted range against the actual gross costs for the years ranging from 2008 to 2014.

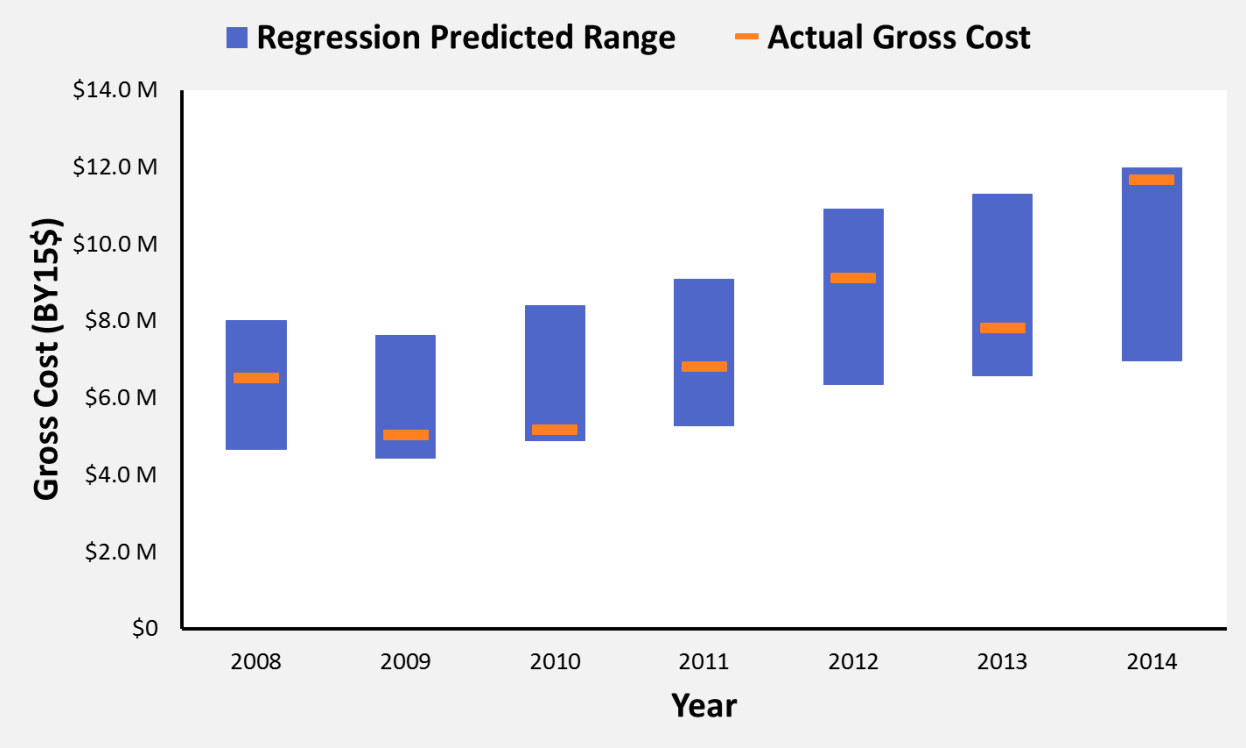

Figure 21. Regression Predicted Range

8 Stochastic Simulation Model

Opposed to the regression model, the simulation model utilizes the data of all job type classifications within the project scope. With an input of estimated total number of houses, the relationship between the input and each type of construction was analyzed, predicting the number of other types of construction by linear regression and time series estimation. Applying bootstrapping and Monte Carlo method, the cost for each classification would be picked from historical data set based on the number of houses and related constructions to get the predicted total residential construction cost

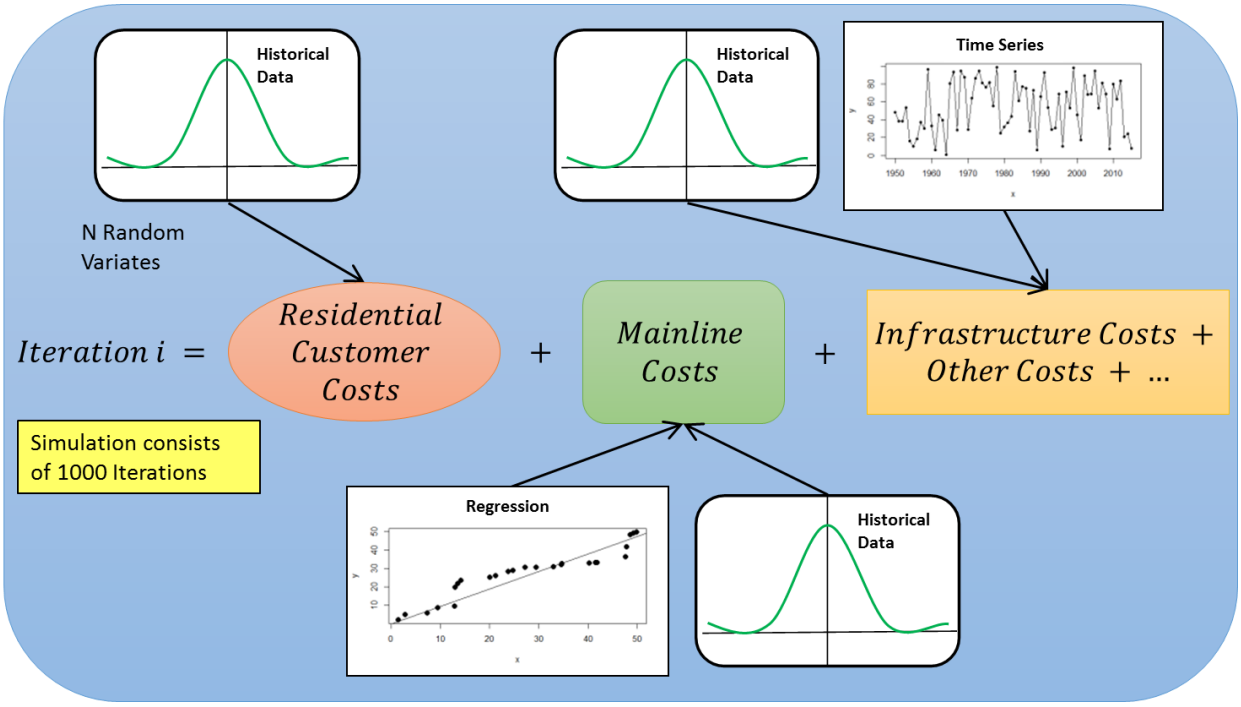

Figure 22. Stochastic Simulation Algorithm

8.1 Model Algorithm

The predicted annual total residential construction cost is divided into two sections. The first part is the direct cost for home constructions. To calculate the cost for the home types it is necessary to first estimate the number of each type of home. The simulation model does this by leveraging the historical data set. For each iteration the breakout by home type is sampled from a distribution of one of the seven years. This breakout is applied to the predicted number of residential customers input to estimate that iteration's breakout of home types.

House Type Percentage							
Year	2008	2009	2010	2011	2012	2013	2014
SFH	77.51%	74.20%	78.20%	71.80%	62.40%	61.59%	60.23%
Condo	0.35%	1.84%	1.86%	0.79%	3.68%	2.09%	5.37%
TH	22.14%	23.96%	19.93%	27.41%	33.93%	36.31%	34.40%

Figure 23. Historical Percentage of Meter Counts for Home Type Jobs

Next, the cost for each type of house will be picked randomly from the corresponding historical data set of cost. Each data set contains the construction cost per house for all the houses of that specific type constructed from 2008 to 2014. The cost will be picked with replacement based on the number of houses in that category. The summation of the construction cost for the three types of houses becomes the residential customer costs.

The second term for the predicted annual total residential construction cost is the cost for the ancillary projects related to house construction. As the number of houses does not have a direct relationship with those related projects, further analysis is needed to predict the number of each other type of construction based on the given number of new residential customers, shown in section 8.2. After determining the number of other related projects described in section 8.2, a similar bootstrapping procedure will be performed as the first part of the predicted annual cost. Each type of other related project has its own historical data sets. This data contains total cost of that project for all the projects completed from 2008 to 2014. The cost will be picked with replacement based on the number of projects in that category, and the summation of all those projects becomes the related costs for house construction.

The summation of the residential customer costs and the related costs for house construction is the predicted annual total residential construction costs for one iteration of the simulation. The simulation will perform 1,000 iterations for each new residential customer to calculate the mean and range for the predicted residential construction costs.

8.2 Job/Project Correlation Analysis

Since there was no existing relationship shown between the number of houses constructed and the number of related projects, regarding as “cost of business” such as barns, garages, etc., further analysis was needed to predict the number of other projects based on the inputted number of houses. As the first observation, Figure 24 demonstrates the correlation among the related other projects to the number of total houses from 2008 to 2014. Based on the results from the table, number of garage projects, infrastructure projects, and other categories projects had negative correlation with the number of houses. Meanwhile, number of barn projects had a correlation of 0.462 to the total number of single family homes; however, the correlation was not able to be considered as good relationship. Thus, according to the existing data and analysis, the team made the assumption that the number of barn projects, garage projects, infrastructure projects, and other categories projects did not have direct relationship to the number of houses inputted. On the other hand, number of mainline projects had a high correlation with the total number of homes in the past 7 years, and the number of mainline projects should be precisely predicted based on the given total number of homes.

Job Classification	Total Number of Homes
Barn	0.310
Garage	-0.483
Infrastructure	-0.0873
Other	-0.314
Mainline	0.952

Figure 24. Correlation between Number of Home and Number of Projects

By plotting the number of mainline projects against number of houses, a linear relationship was demonstrated. Using the linear regression, in order to get the number of mainline projects from number of houses, the coefficient would be 0.033, and the intercept would be -62.005. The R square for this regression is 0.9068. However, there is one thing about this regression that when the total number of house inputted into the model is less than 1,870, the outcome from the regression would be negative. According to the trend of the changes in number of customers in previous, the probability to have a year with less than 2,000 new houses constructed is rare.

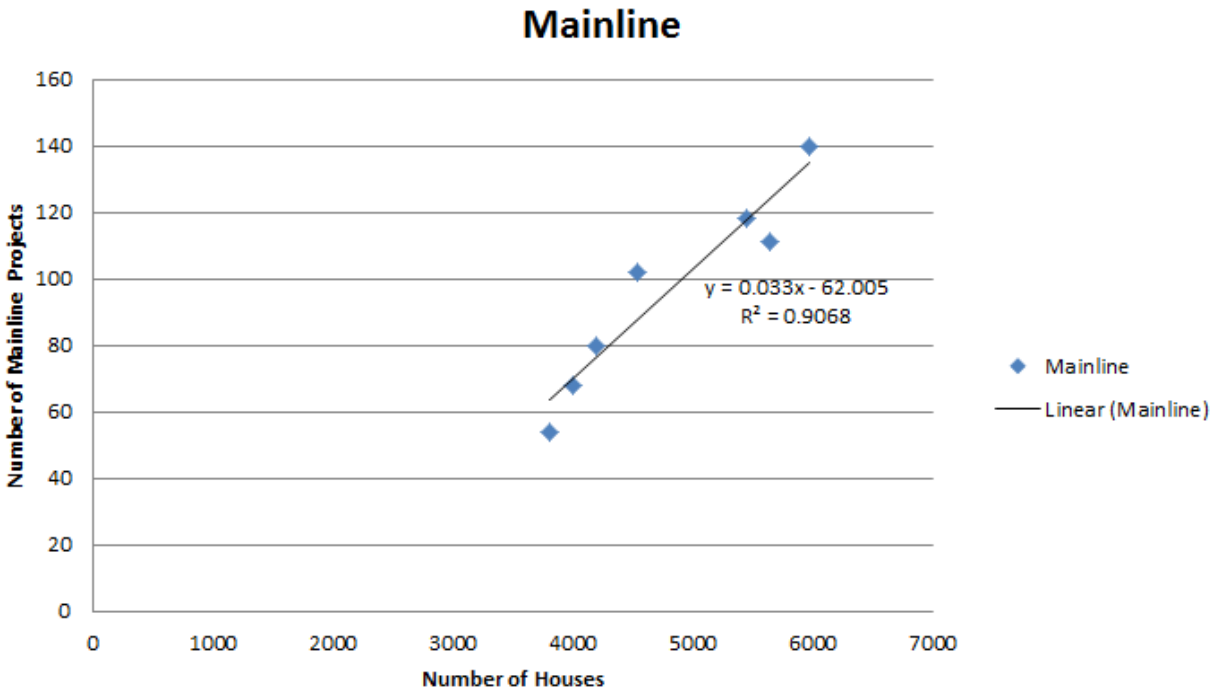

Figure 25. Linear Regression Equation for Predicting Number of Mainline Jobs

Since the other types of related projects did not have a good correlation with the number of house inputted, another method was needed to predict the number of other projects. With a normal development in the recent years, the number of houses and related projects were observed

increasing with a steady rate. As for a short term forecast, that trend should be carried over into next few years. Thus, the time series estimation was one of the best options to predict the number of related projects based on the recent behaviors. As the assumption to apply this method, the number of new customers in next few years should follow the similar trend as they were in the previous years, from 2008 to 2014, and there should not be a huge change in that trend for this short term forecast. In order to capture the entire trend, the ARIMA model was applied to forecast the number of related projects for 2015 to 2018, with an autoregressive of 6 to capture that 7-year period and moving average of 1 to predict number of projects one year ahead. The predicted result is shown in the following table.

Predicted Number of Project				
Year	2015	2016	2017	2018
Barn	8	11	15	17
Garage	14	16	13	12
Infrastructure	19	23	29	10
Other	185	152	118	131

Figure 26. Time Series Prediction for 2015 - 2018

8.3 Simulation Cross Validation

The simulation model was coded in R and the same cross validation technique as the regression model was used to validate our prediction. In the cross-validation, for the prediction of the total cost in each year (e.g. 2008), we omit any corresponding data from our dataset which occurred in that specific year (e.g. all Dwelling and non-dwelling costs for 2008 year). Then, we ran our simulation model and predicted the total cost for that specific year (e.g. 2008) using the new dataset (e.g. 2009-2014). We repeated this procedure for each year and reported the results. In order to show the performance of our simulation model, and compare it to the actual costs, we use two demonstration approaches. First, we use the same table which was used in the regression model in which the average prediction is compared to the actual cost. In the second approach, the results of the 1000 replications of the simulation model were used to build a box plot using 1st quantile, 3rd quantile, median, minimum and maximum of our total cost in those replications.

The following table compares the average prediction of the simulation model to the actual cost in each year. As we can see in the table, in 3 (out of 7 years) we underestimate and in the other 4 years we overestimate, which is pretty consistent with the performance of regression model. The average error of our prediction is 9.42%. The maximum error in the cross-validation technique is 18.3% which occurred in 2010. If we compare the results obtained from simulation model and regression model, we can infer that the performance of simulation model is slightly better. However, we can see the results are pretty much consistent.

Year	Actual Cost (BY15$)	Average Predicted Cost (BY15$)	Percent Error
2008	6.44 M	6.21 M	- 3.56%
2009	4.96 M	5.56 M	+ 12.13%
2010	5.18 M	6.13 M	+ 18.30%
2011	6.74 M	7.06 M	+ 4.85%
2012	9.10 M	8.00 M	- 12.10%
2013	7.29 M	7.84 M	+ 7.54%
2014	10.1 M	9.37 M	- 7.43%

Figure 27. Simulation Cross Validation Results

In the second demonstrating approach, we used the results in 1000 iterations to generate a box-plot. The following diagram shows the box-plots. The red cross shows the actual cost in each year. As the plot shows, in all of the predictions, the actual cost is always between the min and max of the box-plot. Also, in most of the years, the actual cost is either in the box or very close to the box.

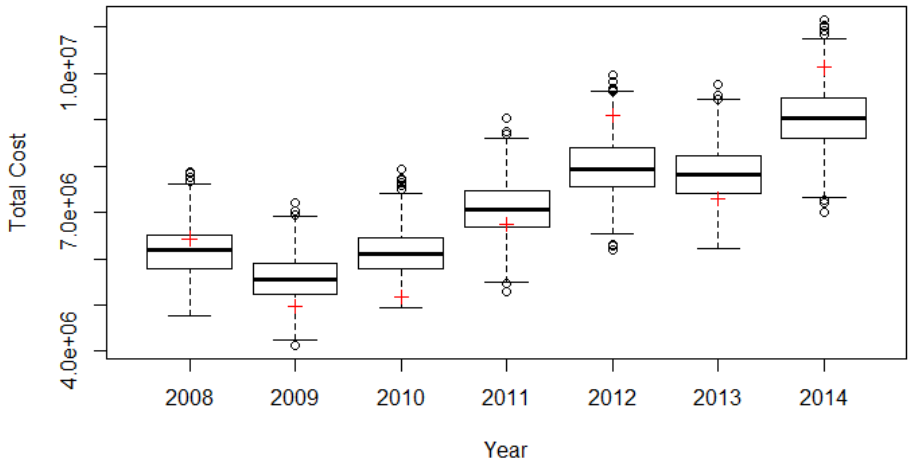

Figure 28. Simulation Predictions vs Actuals

9 Comparison: Regression vs Stochastic Simulation Model

The two different model approaches have been reviewed in detail, the question now is which model should be selected? Should the simulation model be selected because it performs better than the regression model or would it be possible to leverage both models? Figure 29 illustrates a comparison of the forecasted costs from the two different model approaches against the actual gross costs for the years of 2008 to 2014. A trend that is observed is that when one model overestimates, the other model does as well. In other words, both models consistently over and underestimate. For instance, when looking at 2008, both models are underestimating. Additionally, in 2009, both models are overestimating.

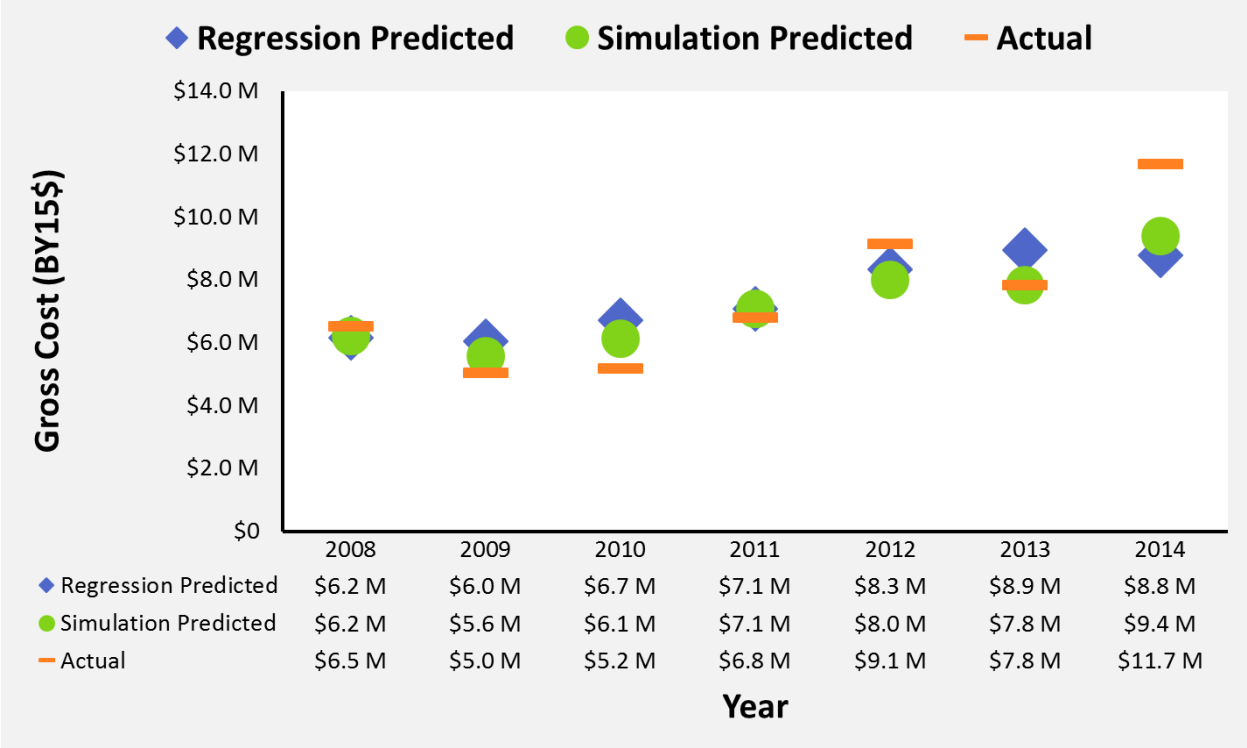

Figure 29. Regression vs Stochastic Simulation Model

Not only does the trend that both models consistently over and underestimate exist, there is another trend that is observed. In cases where both models overestimate, the regression model tends to overestimate more than the simulation model. In instances when both models underestimate, the regression model tends to underestimate more than the simulation model as shown in Figure 30.

	-	+	+	+	-	+	-
Regression	- 6%	+ 20%	+ 29%	+ 4%	- 9%	+ 14	- 25%
Simulation	- 4%	+ 12%	+ 18%	+ 5%	- 12%	+ 8%	- 7%
Year	2008	2009	2010	2011	2012	2013	2014

Figure 30. Comparison of Regression vs Simulation Residuals

As a heuristic by comparing the point predicted values from both approaches for some future year, it can be determined whether they represent over or underestimates. Therefore, if the regression point value is greater than the simulation point value, then the actual cost should be lower than the simulation forecast. Conversely, if the regression point forecast is less than the simulation prediction, then the actual cost should be higher than the simulation.

10 Recommendations

We found this to be both an exciting and challenging project. We found the difference between the types of problems that we have seen in our studies and this real world problem to be very pronounced. In our studies patterns are identifiable, data is extremely clean, and everything we need to know about the problem is in the literature.

We are not saying this problem is completely different from those problems. There were just some challenges that we identified that would make performing a similar analysis more fruitful.

10.1 Feasibility of Ancillary Project References to Supported/Supplied Home Types

Our first recommendation is to investigate whether it is feasible to reference, or tie, some of the Mainline, Infrastructure, Barn, and Garage costs to specific home type jobs.

We are careful to word this recommendation this way. We want to make sure that we are not recommending that this is a necessity. Our lack of knowledge of the industry makes it challenging for us to make some of these judgement calls. That is why we suggest understanding how these costs are allocated and if it is feasible to perhaps label these projects with the development they supply/support.

By doing this it would provide a more accurate allocation of burden to any home type jobs. This is extremely important if analysis like the regression analysis is being done. In our model we had to artificially allocate those costs based on the assumption that the burden projects of a certain year should belong to the home type jobs in that same year. WE understand that this assumption may be lacking when it comes to large jobs like a Mainline due to the fact that a Mainline job supplying power to developments will more than likely cross years. These are large jobs that require longer construction projects.

We initially asked ourselves if it even made sense to allocate Mainline costs. All homes in a local region are supplied by the same mainline infrastructure. We think a more beneficial technique would be to consider adding a prediction of the mainline jobs that will be constructed as an independent variable in future models. It should be investigated by NOVEC if an accurate prediction of the number of Mainline jobs could be done as the prediction of new customers is made as well. By adding this prediction along with new residential customers you are accounting for roughly 90% of the jobs which correspond to approximately 70% of the cost. We believe a model with these two predictors would be much more accurate.

Infrastructure jobs could probably be much easier to allocate to developments. Large development projects could have infrastructure within a set distance allocated to them. We suggest possibly adding a field in the data to label infrastructure jobs to close developments that are related temporally.

The barn and garage jobs could be allocated to projects that occur close and within a relatively short time frame. If a home is connected to a grid and a barn or garage on that same property is connected as such that it requires a stand-alone meter, they should allocated together similarly to the infrastructure jobs. The field relating the development to the barn or garage should reference the home type job. We understand though that this isn't always the case. Barns and garages are

built on properties years, sometimes decades, from when the homes were connected. In this case we think that allocating them to home jobs is not feasible.

It may be more feasible to expand the scope of the input of the model to be more than just the home type jobs we assumed. If the definition of new customers were expanded to include barns and garages as new customers then these jobs could become part of the analysis data set and not required as a burden. New regressions and adjustments to the simulation model would be required for the new additions to the data set. A new completely new analysis would be very interesting to perform on this expanded data set.

10.2 Incorporate more Consistent Data Recording

Our next recommendation concerns quality control and the accuracy of the data set. During the pre-formatting analysis we noticed several job line items were redundant. We also had to remove some jobs in our data that had incorrect or inadequate data recorded.

We eliminated approximately six Mainline jobs that had meters recorded in their line items. We discussed these jobs with NOVEC and they suggested we remove them due to being incorrectly labeled. These jobs may have been mislabeled Mainline. Mainline jobs represent a small amount of jobs in the data set but make up a large percentage of the cost. Having these jobs in the data set would provide a much more accurate cost.

We also came upon several home jobs that did not contain any meter data. Jobs classified as home type projects that do not contain meters were assumed to be incorrectly categorized or being errors. These jobs were removed from the data set.

We also identified some redundancy in the lower level recorded elements. Removing these would make the data easier to work with.

While dealing with the data could be challenging we also thought the NOVEC does a very good job even keeping this historical data. While we identified some challenges we think that further analysis could be done to really help their business. it is an excellent resource.

10.3 Implementation of a Geographic Information System (GIS)

The implementation of a Geographic Information System (GIS) would greatly benefit NOVEC in the long term. In order to implement this fully, the data must be recorded. A possible means to obtain this data is to have personnel from the maintenance department record this. The concept is that all of NOVEC's assets can be incorporated into the GIS which will provide numerous benefits to the enterprise. This will allow NOVEC to locate all of their assets and to capture how different elements relate to each other. As an example, one of the key obstacles that the team faced stemmed from not knowing which Mainline jobs supported which home type jobs. Having this information captured in a GIS system would streamline any need to perform any form of data analysis. Other benefits would include facilitating any field related maintenance work such as for the purposes of locating certain elements in the field and even extending to the engineering planning department to perform advanced spatial analysis. As a short term solution, it would be adequate to record longitude and latitude points associated with each project.

11 Way Forward

For future development in this topic, there are several things that warrant further research and analysis. First and foremost, it would be an important to identify the lead and lagging indicators for the linking of development construction and mainline and infrastructures jobs. Normally, mainline projects and infrastructures project will be completed well before the development construction. The new homes in that specific area can directly hook to the mainline to get electricity. The time differences between the completion of mainline and infrastructure projects and house construction could be years. If the lead/lagging indicators could be clearly identified within the historical data set, the relationship could link the mainline and infrastructure project directly, eliminating the needs for allocating costs to baseline job costs by burdening. Thus, the results from both regression and simulation models can be much more accurate and realistic.

The second major issue is the Other Job classification. As shown in this analysis, that classification accounts for 7% of the total jobs, but 28% in total gross cost. The Other Job classification is a significant source of cost. It is unclear however, what specific jobs are included in that classification. With a further digging into the Other Job classification, the projects that do not fit in the scope of this study should be removed and the short term forecast will be more accurate and reasonable. Further analysis into the "Other" classification may even reveal elements that should be given their own classification for tracking.

Besides the cost analysis and prediction based on the current information, adding more variables such as the effects of terrain or need for trenching may make the forecast more accurate and provide more useful information. In this version of the models, we do not account the differences between the costs of overhead construction and underground construction. Also, having construction in various terrain conditions such as plain or hills could have different costs. In order to do further analysis into these topics, the assistance from the geographic information system would be needed to provide location information.

Last but not least, during the historical data analysis, we often observed that the projects with similar technical attributes tended to have huge cost variances. This was especially true for those dwelling projects with a low number of meters. It could make the short term prediction more accurate if the reasons behind the cost variances could be discovered and accounted for in the models.

12 References

[1] Bowerman, Bruce L., and Richard T. O'Connell. Linear statistical models: An applied approach. Boston: PWS-Kent, 1990.

[2] Draper, Norman, and H. A. R. R. Y. Smith. "Applied regression analysis. Series in probability and mathematical statistics." (1981).

[3] Armstrong, J. Scott (2012). "Illusions in Regression Analysis". International Journal of Forecasting (forthcoming) 28 (3): 689.

[4] David A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press (2005)

[5] R. Dennis Cook; Sanford Weisberg Criticism and Influence Analysis in Regression, Sociological Methodology, Vol. 13. (1982), pp. 313–361

[6] David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 26.A

[7] Hilary L. Seal (1967). "The historical development of the Gauss linear model". Biometrika 54 (1/2): 1–24.

[8] Kroese, D. P.; Brereton, T.; Taimre, T.; Botev, Z. I. (2014). "Why the Monte Carlo method is so important today". WIREs Comput Stat 6: 386–392. doi:10.1002/wics.1314

[9] Efron, B, and Tibshirani, R. J. (1993). An introduction to the bootstrap. Monographs on

Statistics and Applied Probability, No. 57. Chapman and Hall, London. 436pp.

[10] Mooney, Christopher Z., Robert D. Duval, and Robert Duval. Bootstrapping: A nonparametric approach to statistical inference. No. 94-95. Sage, 1993.

[11] Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and Prediction (2nd ed.). New York, NY: Holt, Rinehart and Winston.

[12] Snee, R. D. (1977). Validation of regression models: Methods and examples. Technometrics, 19(4), 415-428.

[13] Mosier, C. I. (1951). Problems and designs of cross-validation. Educational and Psychological Measurement, 11, 5-11.

[14] Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. Journal of the American Statistical Association, 79(387), 575-583.

[15] The Handy-Whitman Index of Public Utility Construction Costs. Whitman, Requardt and Associates. No. 181 , 2015.

[16] "About NOVEC." NOVEC. N.p., n.d. Web. 1 Oct. 2015. <https://www.novec.com/About_NOVEC/index.cfm>.

[17] Gaynor, Michael J. "The Fastest-Growing Suburbs of Washington Are in Counties You've Never Heard Of." Washingtonian. N.p., 01 Apr. 2015. Web. 12 Oct. 2015. <http://www.washingtonian.com/blogs/capitalcomment/real-estate/the-fastest-growing-suburbs-of-washington-are-in-counties-youve-never-heard-of.php>.

[18] "Notation for ARIMA Models". Time Series Forecasting System. SAS Institute. Retrieved 19 May 2015.

[19] Hyndman, Rob J; Athanasopoulos, George. "8.9 Seasonal ARIMA models". Forecasting: principles and practice. oTexts. Retrieved 19 May 2015.

[20] "U.S. Bureau of Labor Statistics." *U.S. Bureau of Labor Statistics*. U.S. Bureau of Labor Statistics, n.d. Web. 15 Sept. 2015. <http://www.bls.gov/>.

13 Appendix A: Project Plan

ID	Task Name	Duration	Start	Finish
1	1 Team Organization and Project Preparation	6 days	Thu 9/3/15	Thu 9/10/15
2	2 Client Meeting	0 days	Thu 9/10/15	Thu 9/10/15
3	3 External Research	7 days	Thu 9/10/15	Fri 9/18/15
4	4 Data Scrubbing	13 days	Fri 9/11/15	Tue 9/29/15
5	5 Data Normalization	21 days	Fri 9/11/15	Fri 10/9/15
6	6 Data Analysis	7 days	Sat 10/10/15	Sat 10/17/15
7	**7 Forecast Model**	**29 days**	**Sun 10/18/15**	**Wed 11/25/15**
8	**7.1 Simulation Model**	**29 days**	**Sun 10/18/15**	**Wed 11/25/15**
9	7.1.1 Model Development	23 days	Sun 10/18/15	Tue 11/17/15
10	7.1.2 Model Testing and Validation	6 days	Wed 11/18/15	Wed 11/25/15
11	**7.2 Regression Model**	**29 days**	**Sun 10/18/15**	**Wed 11/25/15**
12	7.2.1 Alogrithm Establishment	18 days	Sun 10/18/15	Tue 11/10/15
13	7.2.2 Model Testing and Validation	10 days	Thu 11/12/15	Wed 11/25/15
14	**8 Deliverables Preparation**	**12 days**	**Wed 11/25/15**	**Thu 12/10/15**
15	8.1 Final Report	12 days	Wed 11/25/15	Thu 12/10/15
16	8.2 Final Presentation	12 days	Wed 11/25/15	Thu 12/10/15
17	8.3 Website	5 days	Wed 12/2/15	Tue 12/8/15
18	9 On Site Presentation	0 days	Fri 12/4/15	Fri 12/4/15
19	10 Final Presentation Due	0 days	Fri 12/11/15	Fri 12/11/15
20	11 Final Report Due	0 days	Fri 12/11/15	Fri 12/11/15

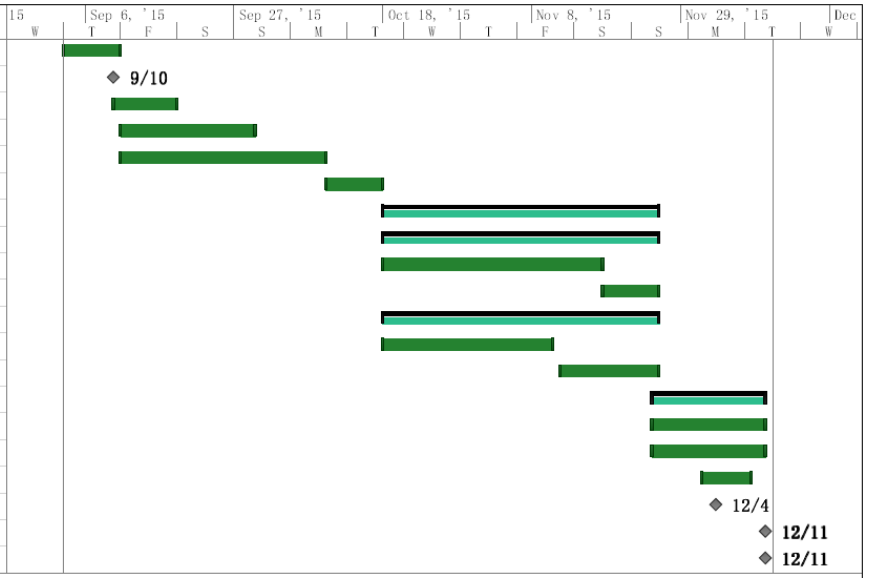

Figure 31. Project Plan

Throughout the entire project, the three-month development period was divided into three major sections. Starting from September 3, 2015, the team conducted activities to understand the project and dig into the data set given by the client. The team gained background information about NOVEC and the project from the initial meeting with the client before doing a series of research and literature reviews to determine the methods and techniques that would apply in the analysis and forecasting of the project. After receiving the data set from the client, the team spent more than one month to scrub the data set to an analyzable format as well as normalized the data to the same cost level of 2015 dollars by calculating and applying inflation factors to each year.

After selecting the suitable portion of the data set, the team built two forecast models from different perspectives within five weeks in the second major section of the development. The team was divided into two groups to create and valid the model they built. Several meetings with professors were conducted during this period for assistance. During this period, each group spent great amount of time to set the algorithm of the prediction, then used cross validation to test the results as well as comparing the results between two models.

In the final three weeks, the focus transferred to prepare the deliverables after the completion of forecast models. The team kept updating the content of presentation and report, making modifications based on the suggestions from professors and classmates. The team also visited NOVEC facility to present the project to the client, gaining feedback of the progress, before the final presentation on December 11, 2015.

14 Appendix B: Data Dictionary

WR_NO – Work Request Number. This is a unique identifier for the individual construction job that was performed.

WR_TYPE_CODE – Work Request Type Code. NOVEC's description of the type of job.

WR_TYPE_DESC – A description of WR_TYPE_CODE

COMPLETE_DATE – The day the job was completed. We are ignoring length of construction for this project.

WORK_CATEGORY_TYPE_CODE – Categorization of work for accounting purposes (e.g., new construction, replacement, etc.)

WORK_CATEGORY_TYPE_DESC – A description of WORK_CATEGORY_TYPE_CODE

CLASSIFICATION_CODE – This is the type of structure we are providing service to.

CLASSIFICATION_DESC – A description of CLASSIFICATION_CODE

RUS_CODE – Classification of the Type of Service. "0100" is Overhead construction, "0101" is Underground construction.

CU_ID – Construction Unit ID. This is the actual component being built/installed.

UOM – Unit of Measure. What type of unit does the Quantity represent.

CU_TYPE_CODE – A categorization of the type of Construction Unit. Broad grouping based upon general use of Construction Unit.

CU_CATEGORY_CODE – A further category refinement of the type of Construction Unit. Broad grouping based on type of Construction Unit.

CU_CATALOG_ID – A further category refinement of the type of Construction Unit. More refined grouping based on type of Construction Unit.

ACTION_CODE – What action did we take during construction. I – Install, R – Remove, T – Transfer (take from old installation and reuse in new), A – Abandon

QTY – Quantity. How many of the Construction Units were acted upon (installed/removed/transferred/abandoned)

LABOR_COST – Total labor cost for the Construction Unit quantity and action

MATERIAL_COST - Total material cost for the Construction Unit quantity and action

OVERHEAD_COST - Total overhead cost for the Construction Unit quantity and action

CONTRIBUTION – How much (if any) the customer was required to pay NOVEC towards the cost of construction

The Construction Unit categorization works like this: CU_TYPE is a broad grouping of the unit (typically Overhead or Underground) based on the type of construction it is primarily used in. CU_CATEGORY is

the most general classification of the type of unit (e.g., POLE, TRANSFORMER, CONDUCTOR, etc.). CU_CATALOG is then a breakdown/refinement of the type of unit within CU_CATEGORY

15 Appendix C: Refined Data Vector Examples

WR_NO	COMPLETE_DATE	BARN	CONDO	GARAGE	MAINLINE	MOBILE HOM	OTHER	SGL FAMILY HOME	TOWNHOUSE	INFRASTUCTUR	RUS_CODE	METERS	LABOR_COST (BY15$)	MATERIAL_COST (BY15$)	OVERHEAD_COST (BY15$)	CONTRIBUTION (BY15$)	Ft_of_Conductor	Trenching_Machine	Labor_Labor
76302	38414	0	0	0	0	0	1	0	0	0	101	0	36440.03759	117044.4602	25187.28156	0	30764	8592	0
77884	38589	0	0	0	0	0	1	0	0	0	101	0	20643.9525	51803.89954	12381.39925	-27183.5598	38704	16608	108
78360	38392	0	0	0	0	0	0	1	0	0	101	0	0	0	0	-314.642754	0	0	0
78525	38582	0	0	0	0	0	0	1	0	0	101	5	7092.518675	3126.282502	1829.882181	-287.7640374	2464	2164	96
78721	39926	0	0	0	0	0	0	1	0	0	101	2	260.834855	130.736965	109.5827122	-54.25085483	340	60	0
78734	40316	0	0	0	0	0	0	1	0	0	101	2	218.5141672	209.5350927	101.8037139	-54.73576917	540	20	0
78948	39821	0	0	0	0	0	0	1	0	0	101	2	300.2101254	68.05986742	117.883744	-59.67594032	120	160	0
79527	38742	0	0	0	0	0	0	1	0	0	101	2	408.5992876	117.7781159	150.7024268	-50.0174002	220	220	0
81765	38385	0	0	0	0	0	1	0	0	0	101	0	6465.137076	50967.73347	8014.659311	0	29608	0	0
82896	38400	0	0	0	0	0	1	0	0	0	101	0	136655.6478	167647.7802	66412.26735	-89964.37099	200866	73330	54
83568	38839	0	0	0	1	0	0	0	0	0	101	0	0	4974.213302	497.423045	0	0	0	0
83582	38839	0	0	0	1	0	0	0	0	0	101	0	0	14955.37415	1495.537415	0	0	0	0

Figure 32. Refined Data Vector Examples

16 Appendix D: Other Regressions

16.1 Log Transformed and Log Squared Regression Analysis

Another form of regression we investigated was log transformed regressions of our data set. This type of analysis requires transforming the data in log space performing a linear regression and transforming it back. The procedure is:

$$log(y) = a * log(x) + b$$

$$exp(log(y)) = exp(a * log(x) + b)$$

$$y = exp(b) * x^a \qquad \leftarrow \text{This is the final form of the equation}$$

We reran all of our regressions with the log transformed in the hope that we could improve our predictions of the actual costs. This analysis could not be calculated without a value for the intercept. This reduced the number of regression we had to only three per parameter.

16.1.1 2005 – 2015 Data Set

16.1.1.1 Log Transformed Results - Machine Trenching

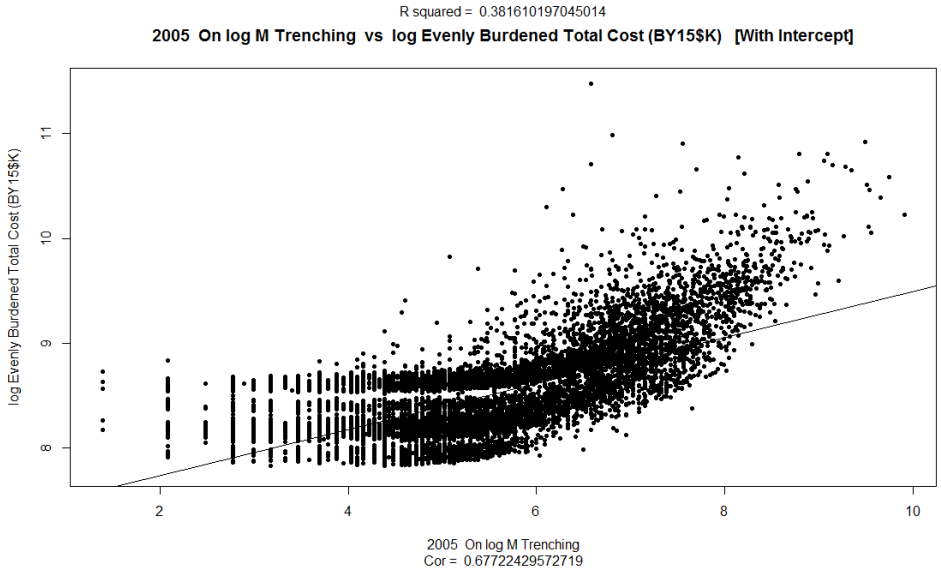

Figure 33. Log Transformed Results: Gross Cost with Burden (Equally Distributed) vs Machine Trenching

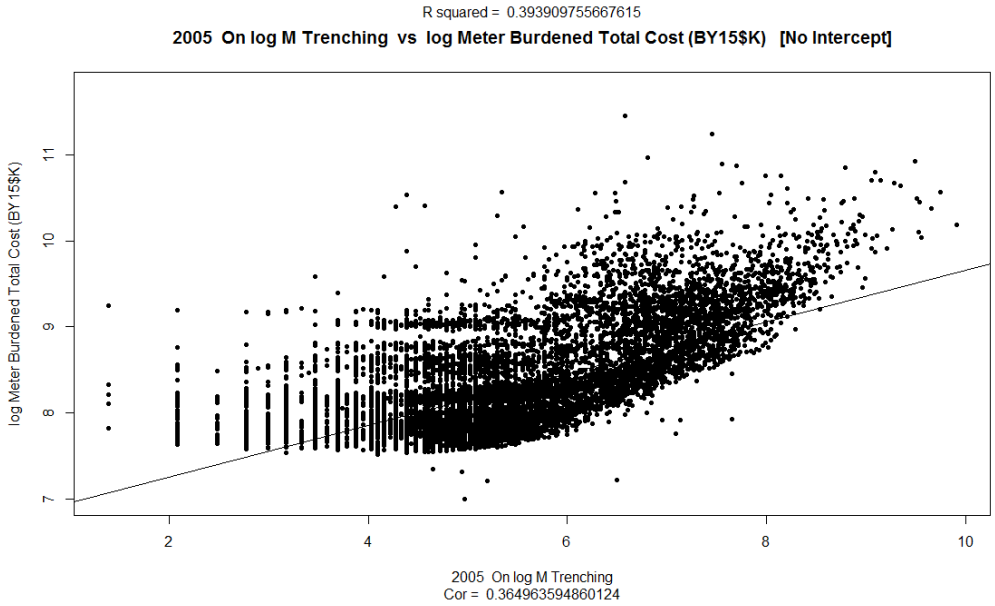

Figure 34. Log Transformed Results: Gross Cost with Burden (Meter) vs Machine Trenching

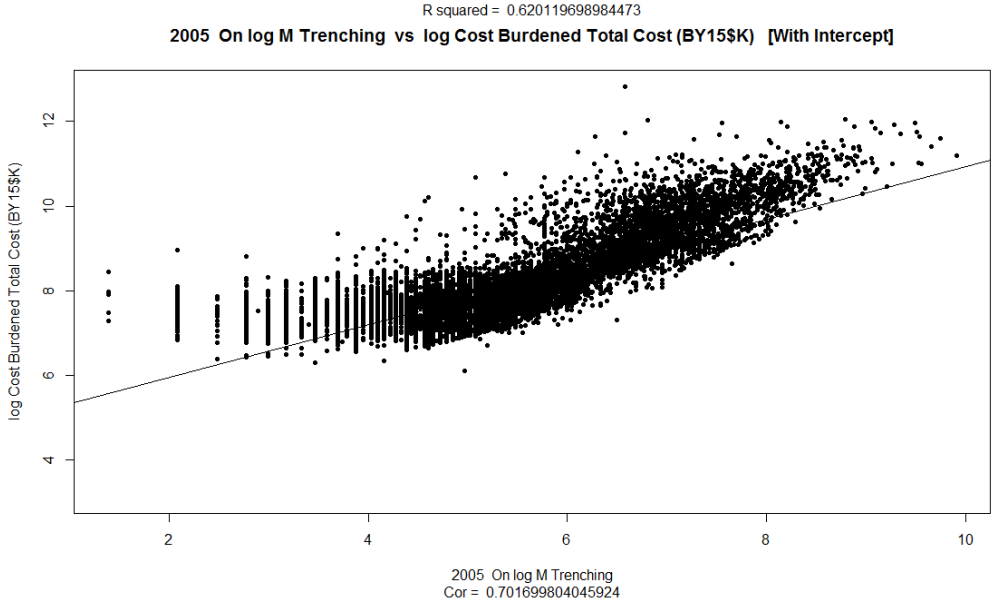

Figure 35. Log Transformed Results: Gross Cost with Burden (Gross Cost) vs Machine Trenching

16.1.1.2 Log Transformed Results - Labor

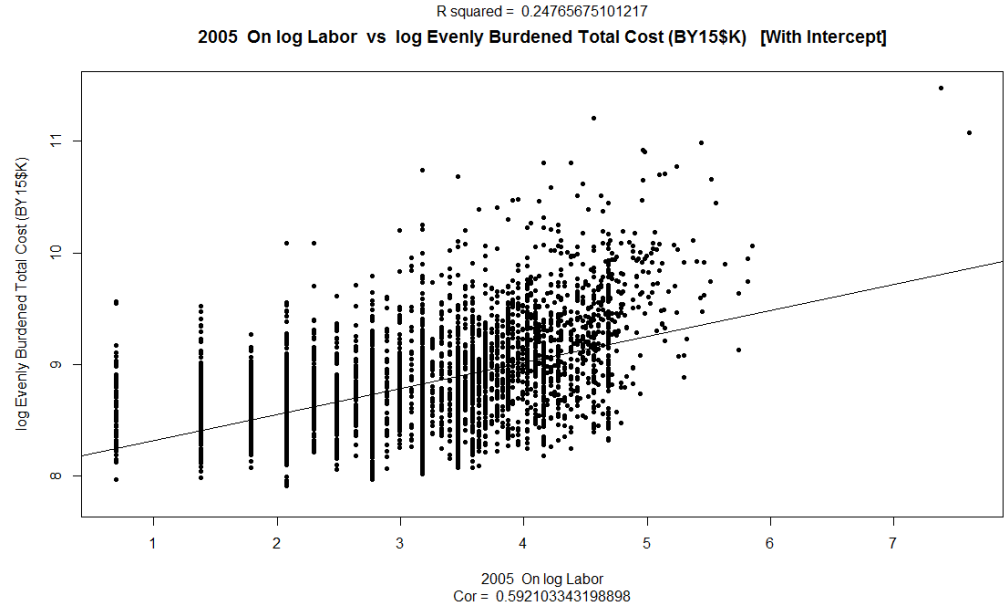

Figure 36. Log Transformed Results: Gross Cost with Burden (Equally Distributed) vs Labor

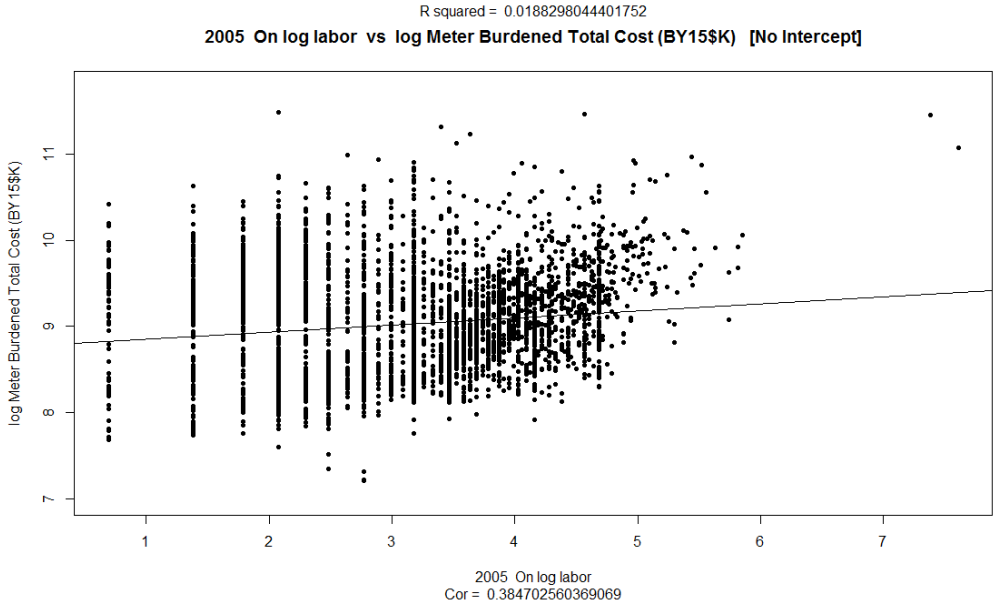

Figure 37. Log Transformed Results: Gross Cost with Burden (Meters) vs Labor

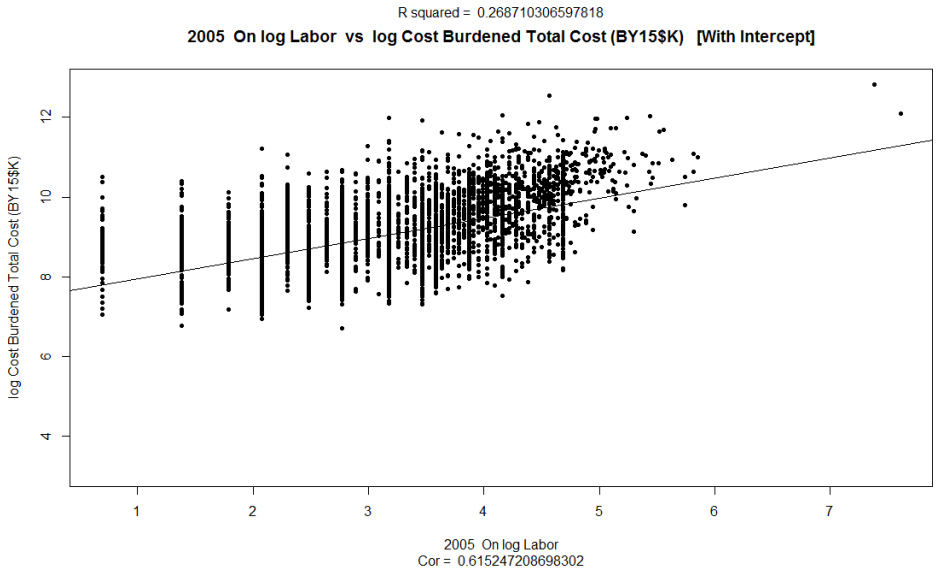

Figure 38. Log Transformed Results: Gross Cost with Burden (Gross Cost) vs Labor

16.1.1.3 Log Transformed Results - Transformers

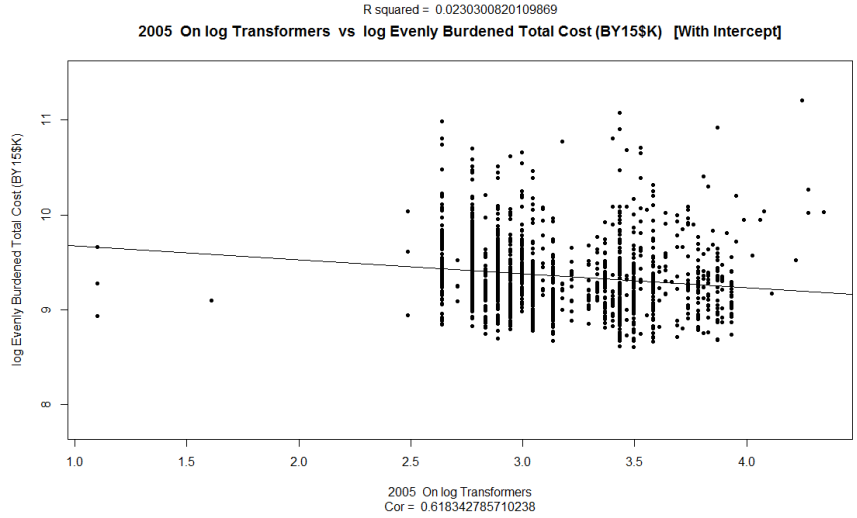

Figure 39. Log Transformed Results: Gross Cost with Burden (Equally Distributed) vs Transformers

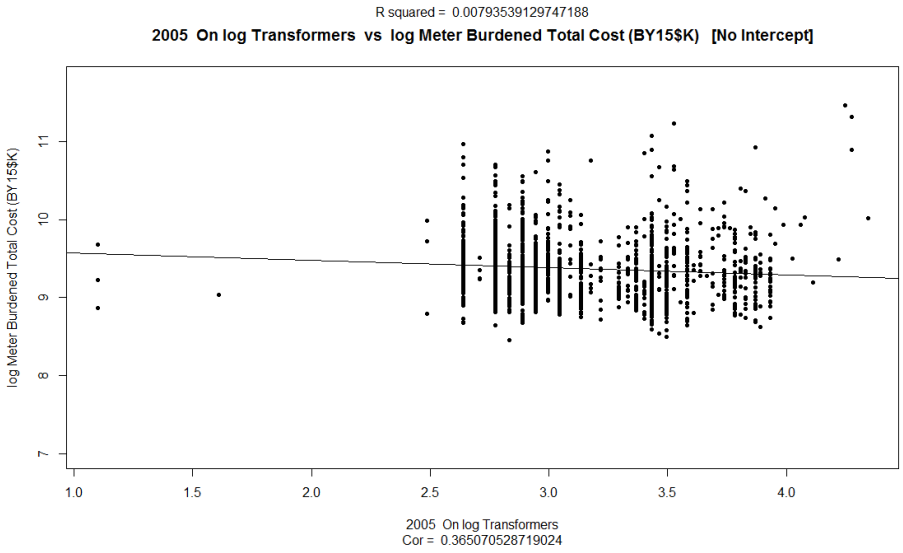

Figure 40. Log Transformed Results: Gross Cost with Burden (Meters) vs Transformers

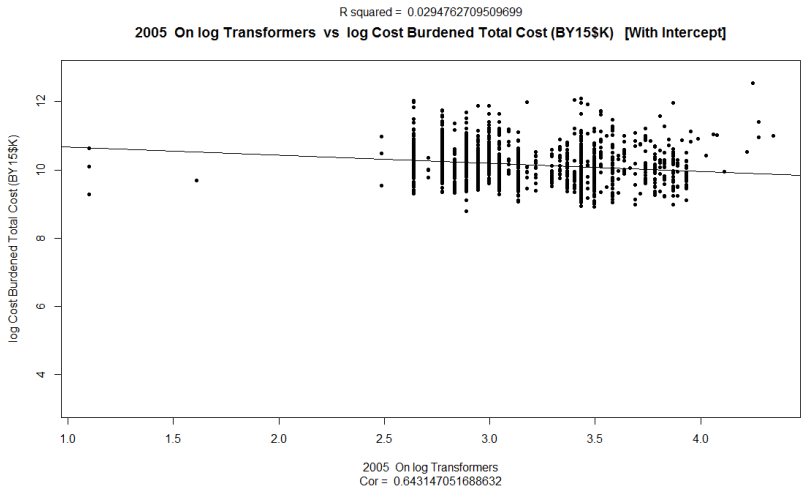

Figure 41. Log Transformed Results: Gross Cost with Burden (Gross Cost) vs Transformers

16.1.2 2008 – 2015 Data Set

16.1.2.1 Log Transformed Analysis - Number of Meters

We began with the regression with number of meters as a parameter. The regressions evaluated each type of burden. We analyze the regressions versus the plots of the data. The R squared and correlation in the data set was calculated but this concerned the regressions prediction of jobs to cost. We have to test the data by taking the actual meter counts and transforming them to jobs based on historical data. We can then compare the costs projected to the actual costs incurred.

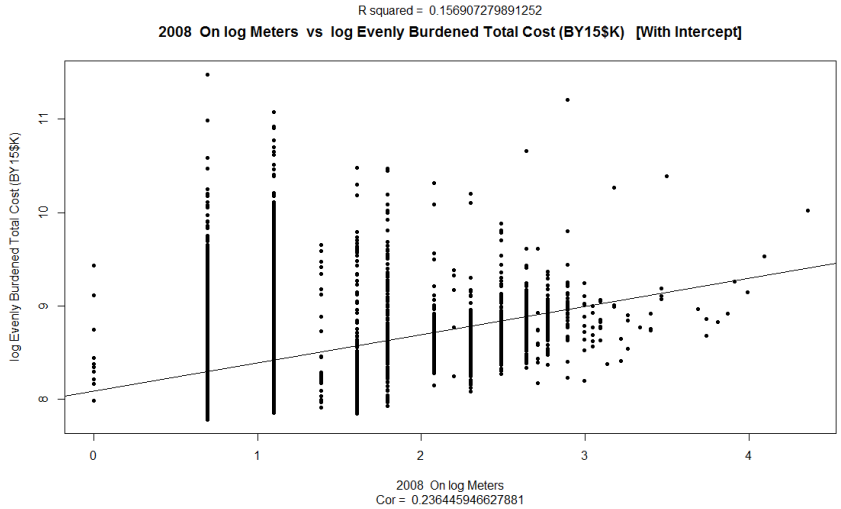

Figure 42. Log Transformed: Gross Cost with Burden (Equally Distributed) vs Number of Meters

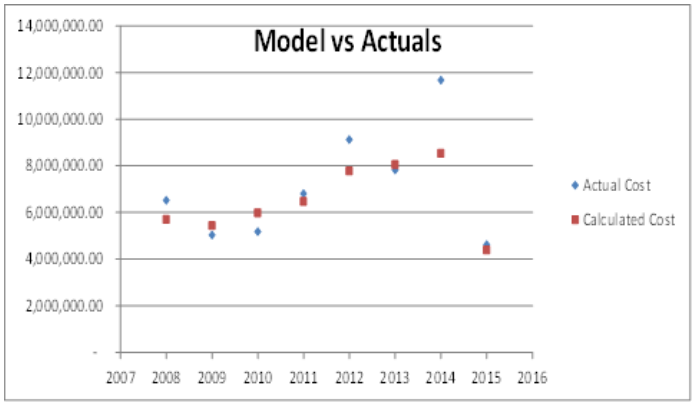

Figure 43. Log Transformed: Gross Cost with Burden (Equally Distributed) predicted from Number of Meters Model against Actual Gross Costs

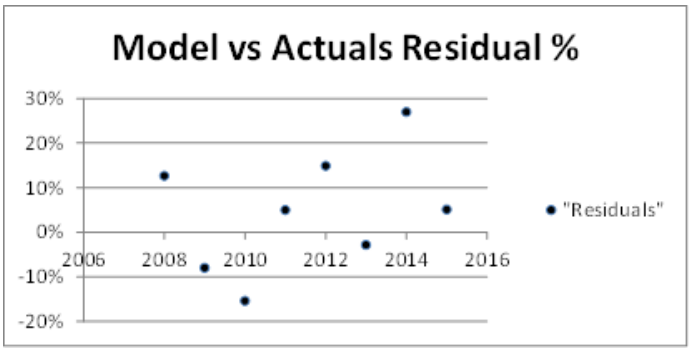

Figure 44. Log Transformed: Gross Cost with Burden (Equally Distributed) predicted from Number of Meters Model Residuals

The chart above shows the meter counts does not make a very good linear predictor of cost in log space. Again we see a familiar problem occur. We have jobs with equal meter counts but with extreme variance in the cost. This makes prediction challenging.

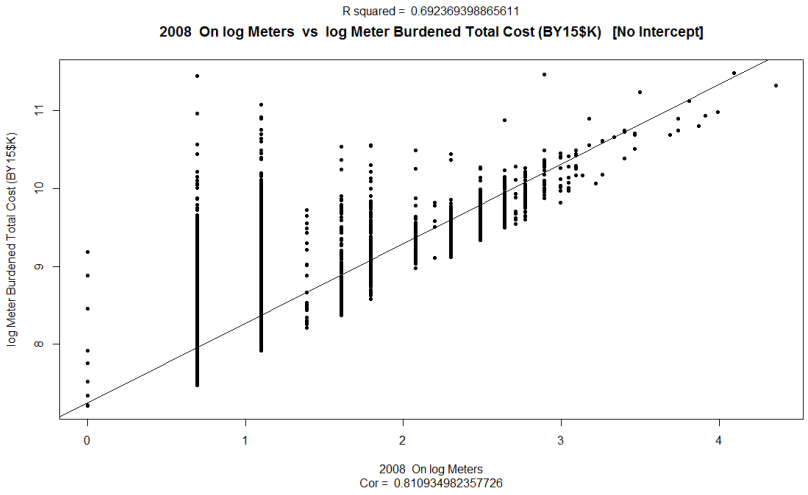

Figure 45. Log Transformed: Gross Cost with Burden (Meter) vs Number of Meters

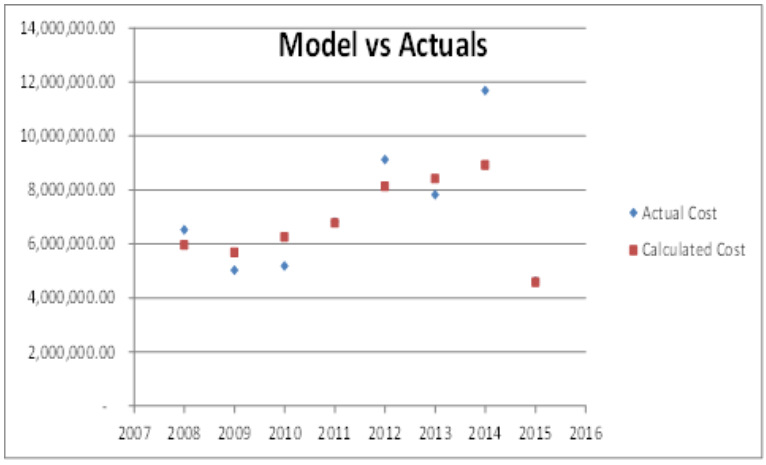

Figure 46. Log Transformed: Gross Cost with Burden (Meter) predicted from Number of Meters Model against Actual Gross Costs

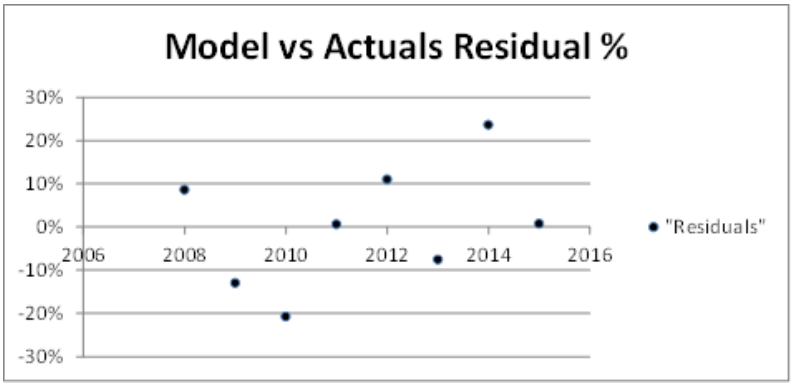

Figure 47. Log Transformed: Gross Cost with Burden (Meter) predicted from Number of Meters Model Residuals

We see that transforming the data into log space does benefit the meter allocated burden costs. The variance in costs for similar meter counts is still present but it is much less pronounced in the high end of the domain of the data set. At the low end of the domain it would not be a very good predictor.

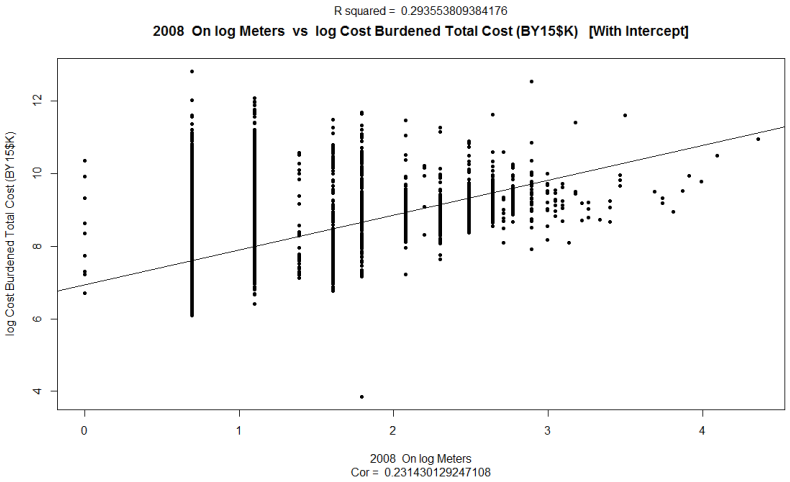

Figure 48. Log Transformed: Gross Cost with Burden (Gross Cost) vs Number of Meters

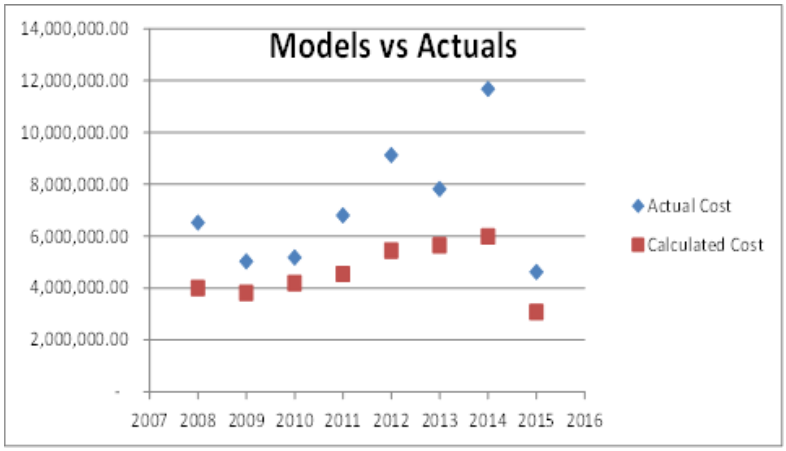

Figure 49. Log Transformed: Gross Cost with Burden (Gross Cost) predicted from Number of Meters Model against Actual Gross Costs

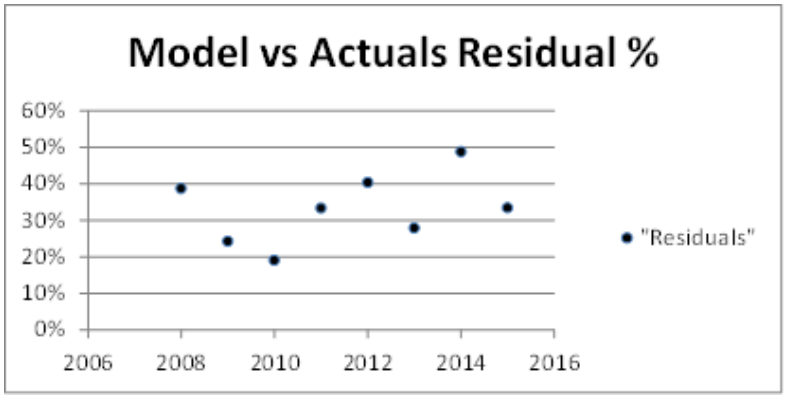

Figure 50. Log Transformed: Gross Cost with Burden (Gross Cost) predicted from Number of Meters Model Residuals

The cost allocated burden cost is similar to the evenly allocated burden cost. The variance along the range somewhat invalidates meters as very good predictor in log space. The cost allocated burden consistently underestimated the actuals by a relatively high magnitude. This is an inadequate model.

16.1.2.2 Log Transformed Analysis - Feet of Conductor

Next we evaluated the log transformed Feet of Conductor as a predictor of cost. As expected the transformation created a more linear placement of the data. We hoped this would lead to a better projected cost than the unit space models.

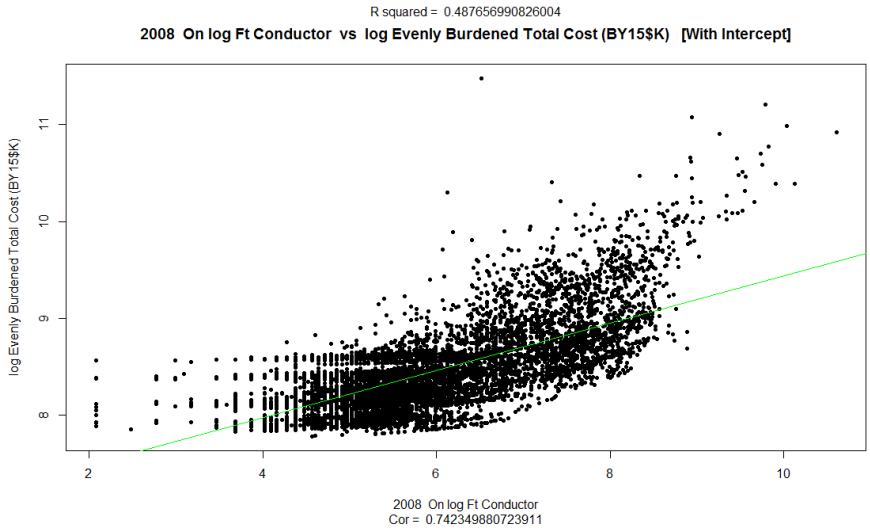

Figure 51. Log Transformed: Gross Cost with Burden (Equally Distributed) vs Feet of Conductor

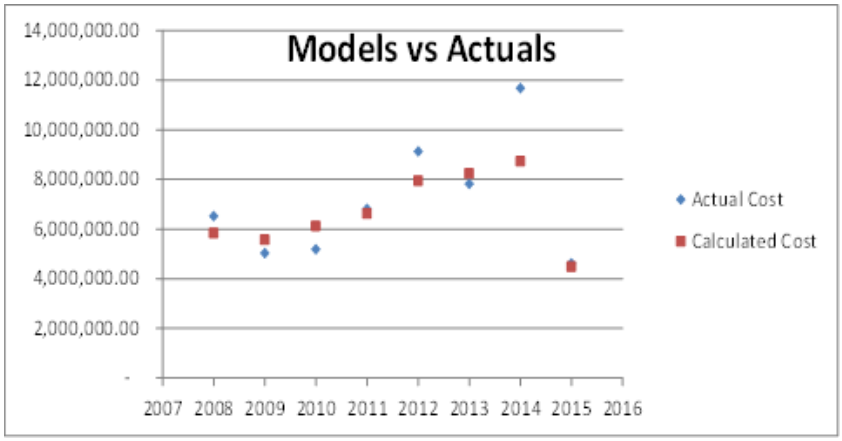

Figure 52. Log Transformed: Gross Cost with Burden (Equally Distributed) predicted from Feet of Conductor Model against Actual Gross Costs

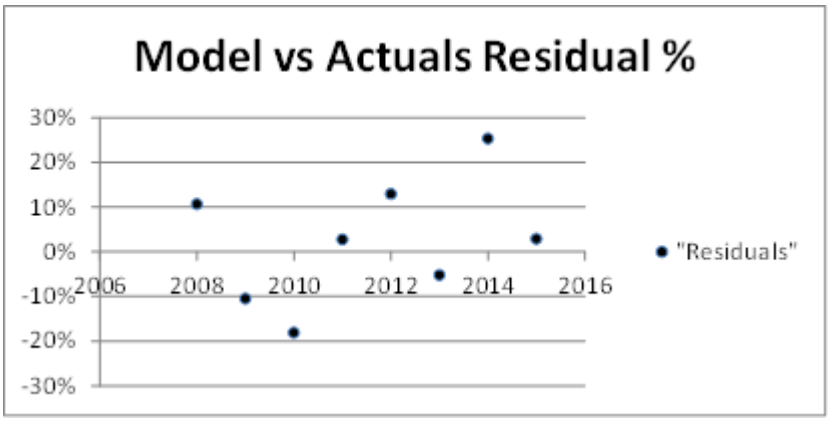

Figure 53. Log Transformed: Gross Cost with Burden (Equally Distributed) predicted from Feet of Conductor Model Residuals

The evenly burdened cost model did not adequately model the actuals. We also noticed that the log transformed even data seemed to follow a quadratic trend. We explored this regression later. The explanatory power of these models was relatively good as the unit space models but not terrific.

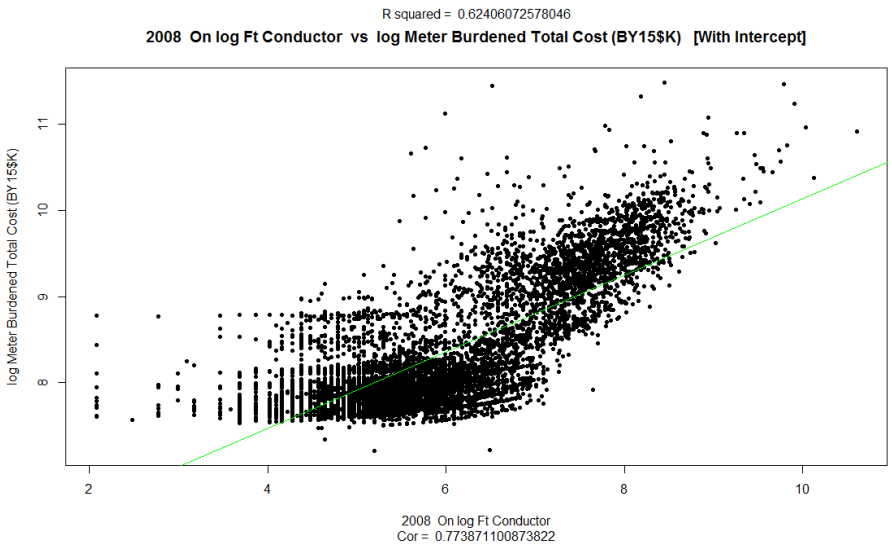

Figure 54. Log Transformed: Gross Cost with Burden (Meter) vs Feet of Conductor

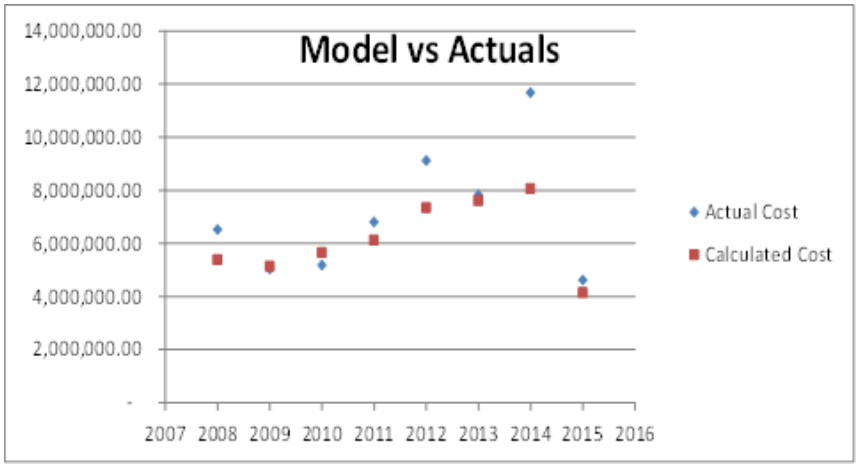

Figure 55. Log Transformed: Gross Cost with Burden (Meter) predicted from Feet of Conductor Model against Actual Gross Costs

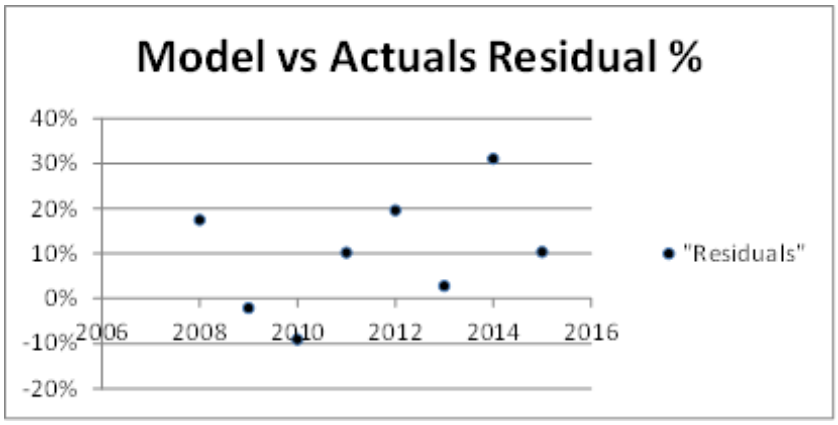

Figure 56. Log Transformed: Gross Cost with Burden (Meter) predicted from Feet of Conductor Model Residuals

The meter burdened cost had relatively higher error then the other models.

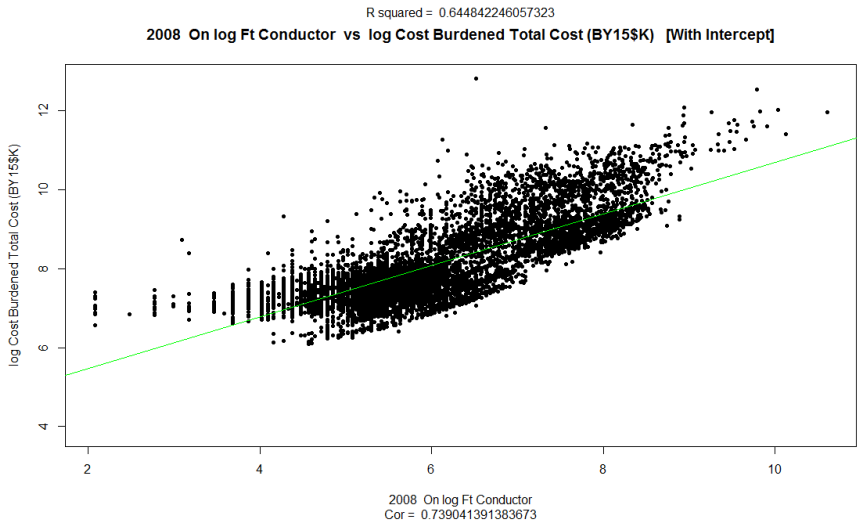

Figure 57. Log Transformed: Gross Cost with Burden (Gross Cost) vs Feet of Conductor

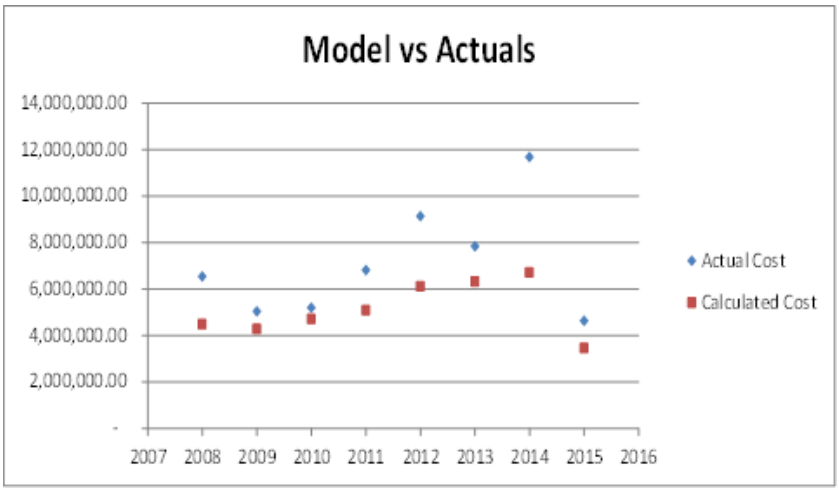

Figure 58. Log Transformed: Gross Cost with Burden (Gross Cost) predicted from Feet of Conductor Model against Actual Gross Costs

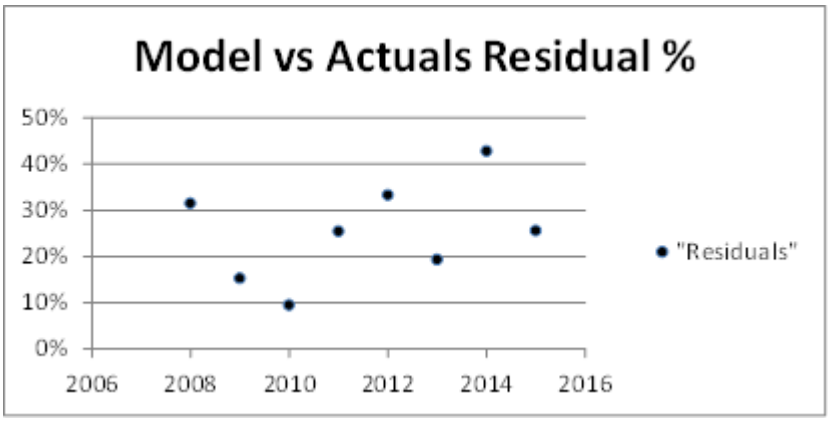

Figure 59. Log Transformed: Gross Cost with Burden (Gross Cost) predicted from Feet of Conductor Model Residuals

We expected this to be a very good model when we looked at the plot. We were disappointed with the higher error in this model. It consistently underestimated the actuals.

16.1.2.3 Log Squared – Feet of Conductor

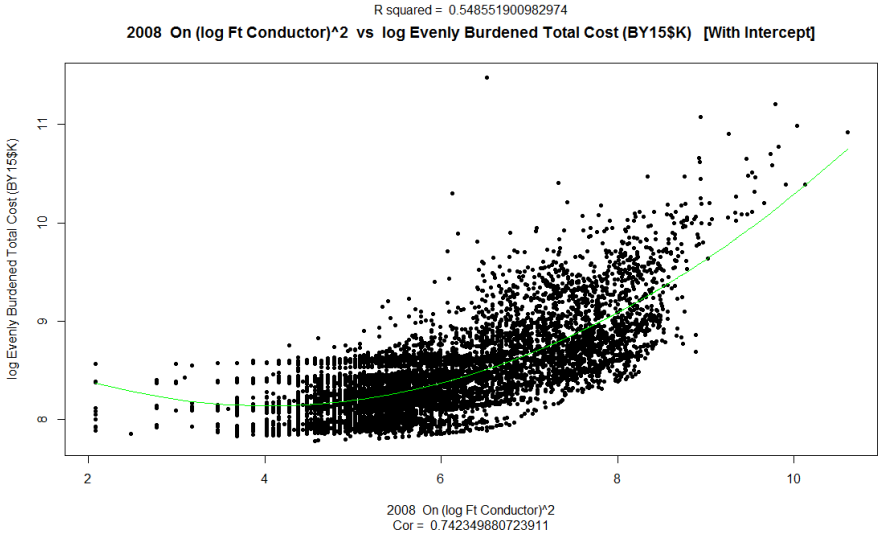

Figure 60. Log Squared: Gross Cost with Burden (Equally Distributed) vs Feet of Conductor

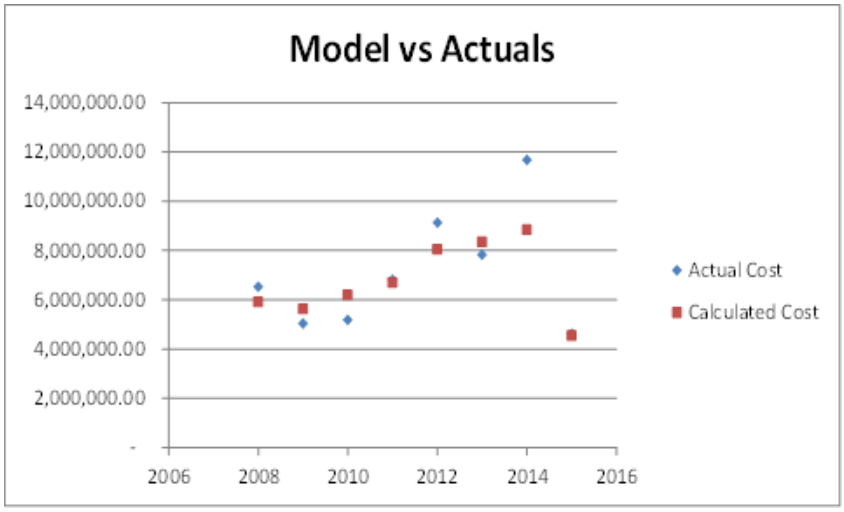

Figure 61. Log Squared: Gross Cost with Burden (Equally Distributed) predicted from Feet of Conductor against Actual Gross Costs

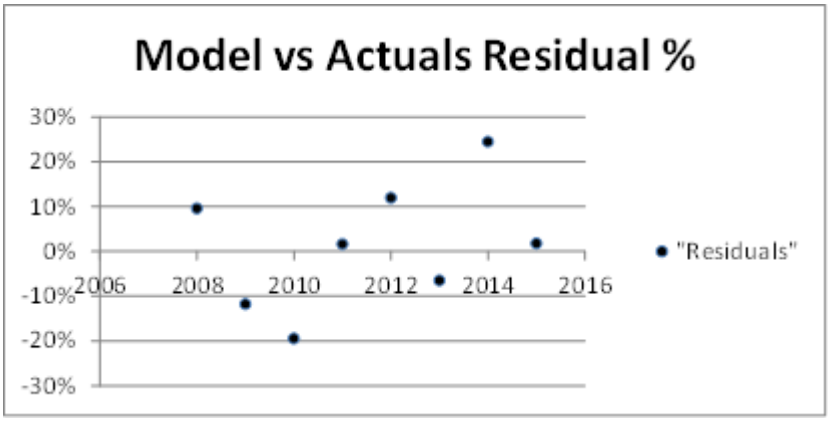

Figure 62. Log Squared: Gross Cost with Burden (Equally Distributed) predicted from Feet of Conductor Residuals

The log squared models were formulated as the regular log transformed models but we squared the log(x) term. The model is of the form:

$$y = exp(a * log(x)\text{^}2 + b)$$

We leveraged the R statistical language to run these regressions as we did the other regressions.

The evenly burdened model was a relatively good model that both overestimated and underestimated the data.

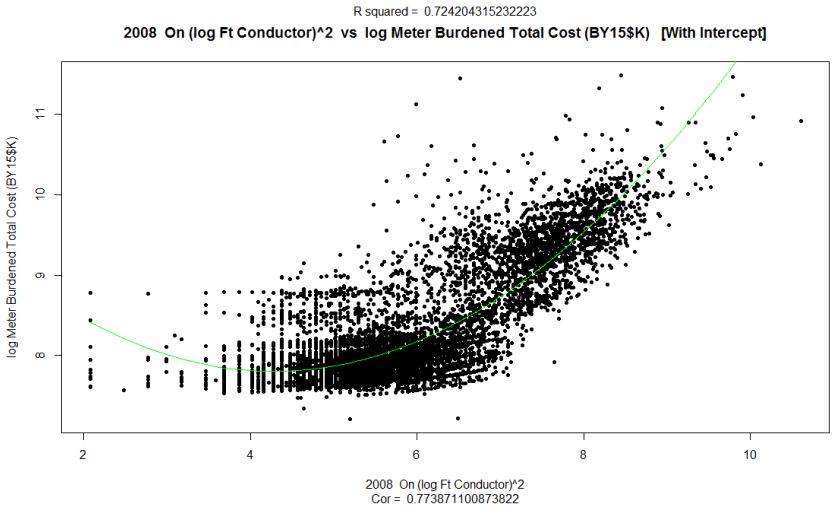

Figure 63. Log Squared: Gross Cost with Burden (Meter) vs Feet of Conductor

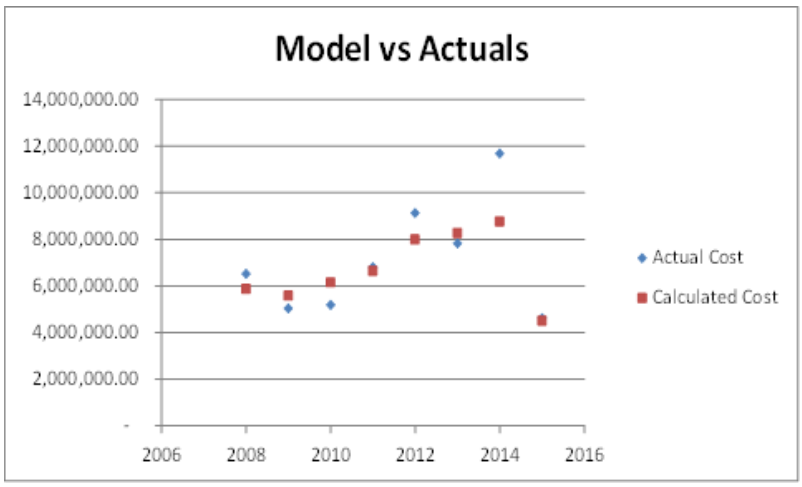

Figure 64. Log Squared: Gross Cost with Burden (Meter) predicted from Feet of Conductor against Actual Gross Costs

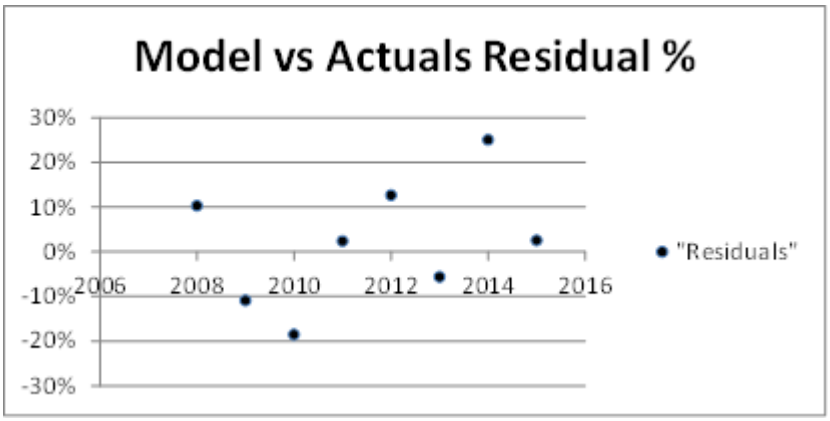

Figure 65. Log Squared: Gross Cost with Burden (Meter) predicted from Feet of Conductor Residuals

The equally distributed burdened model also had middling results. It still did not surpass the better performing unit space models for prediction.

Mean Squared Error	Model
1.68E+12	Model: Log Meter Model Even Burden
1.36E+12	Model: Log Meter Model Meter Burden
8.37E+12	Model: Log Meter Model Cost Burden
1.49E+12	Model: Log Ft Conductor Model Even Burden
2.31E+12	Model: Log Ft Conductor Model Meter Burden
5.72E+12	Model: Log Ft Conductor Model Cost Burden
1.41E+12	Model: Log Ft Conductor Model Even Burden Squared

Figure 66. Comparison of Mean Squared Error across Log Models

Figure 66 above shows the Mean Squared Error of the log space transformed models. The highlighted models are the only log space models that have similar performance to the better performing unit space regressions. None of the log models perform as well. The added complexity of the models do not warrant their use. This lead to our choice of a unit space model

16.2 2005 – 2015 Data Set

16.2.1 Gross Cost with Burden (Equally Distributed)

The chart below plots the adjusted R-squared value on the y-axis and the potential predictor variables on the x-axis for a regression equation to calculate the response variable, gross cost with burden (equally distributed). The graph should be viewed horizontally, and white spaces indicate that a certain predictor variable will not be used. As the graph is viewed horizontally, blocks with colors indicate that those predictors are needed to produce a certain R-squared value. For instance, looking at the lowest level, in order to obtain a regression that has an R-squared value of 0.35 then an intercept is needed as well as the predictor variable, labor. As another example, to generate an R-squared value of 0.68, an intercept is needed and the predictors variables of conductor cable and labor.

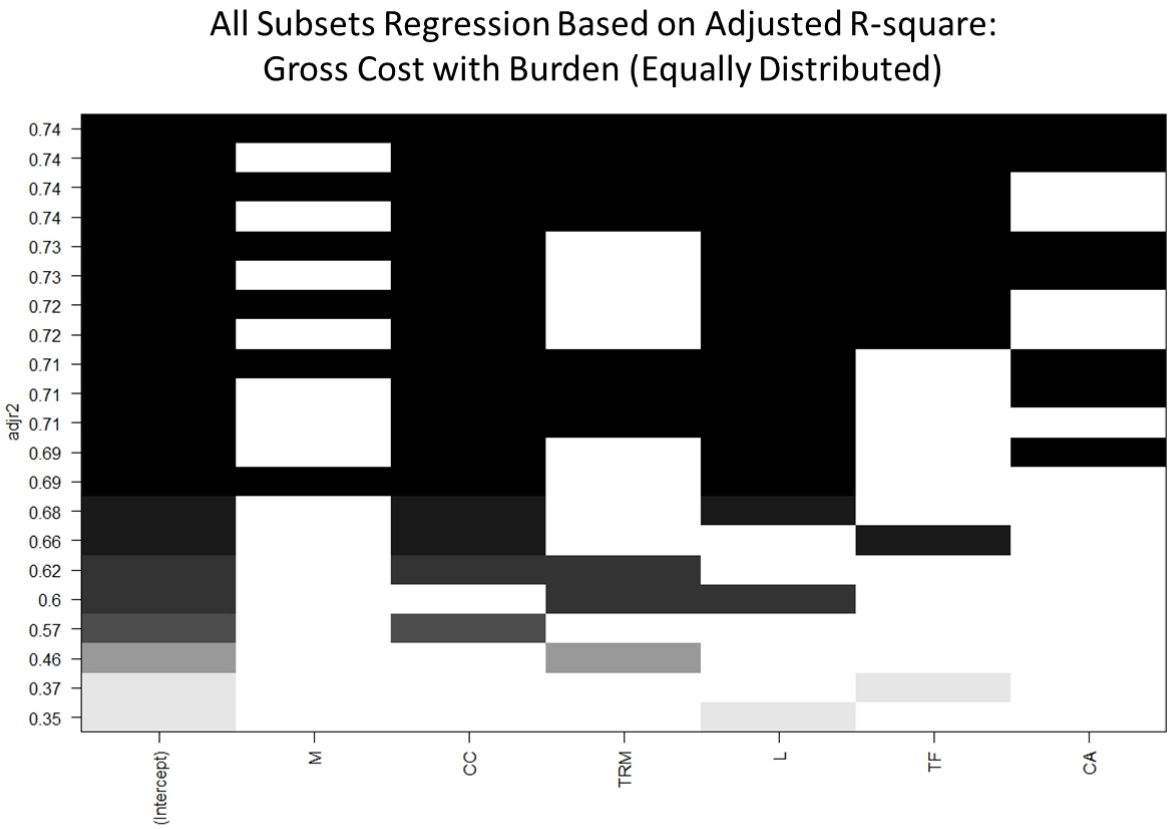

Figure 67. All Subsets Regression Based on Adjusted R-square: Gross Cost with Burden (Equally Distributed)

16.2.1.1 Simple Linear Regression

16.2.1.1.1 Meters

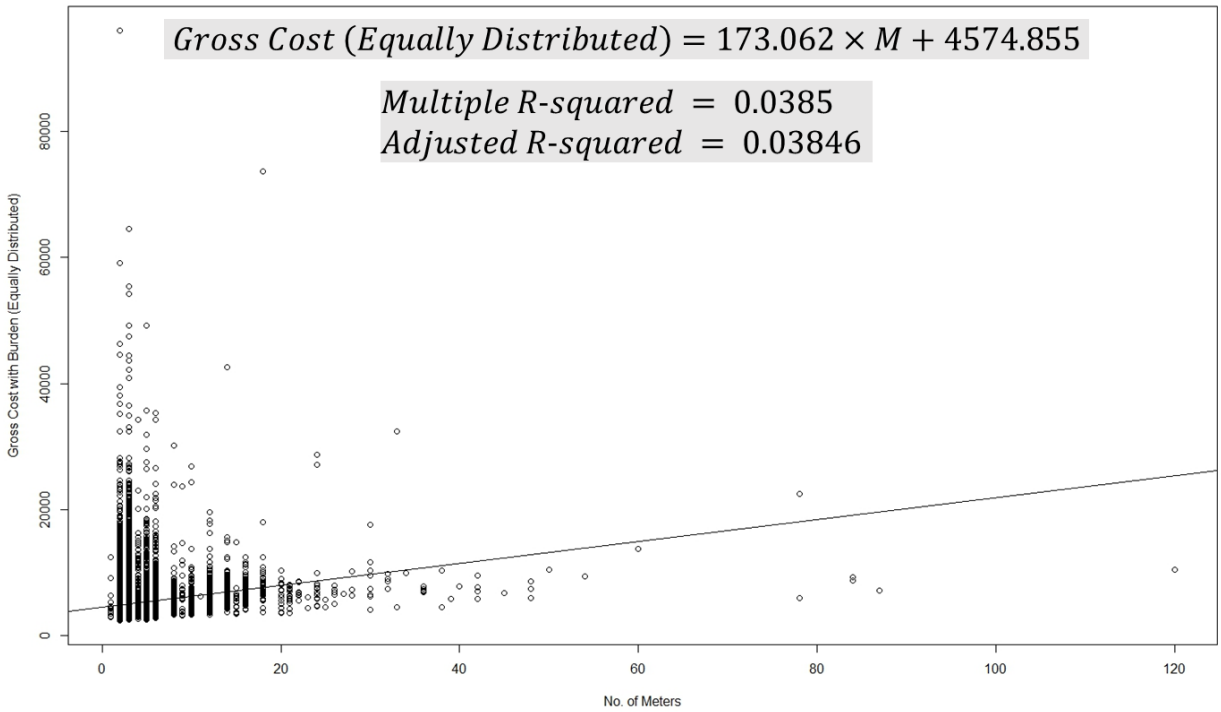

Figure 68. Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

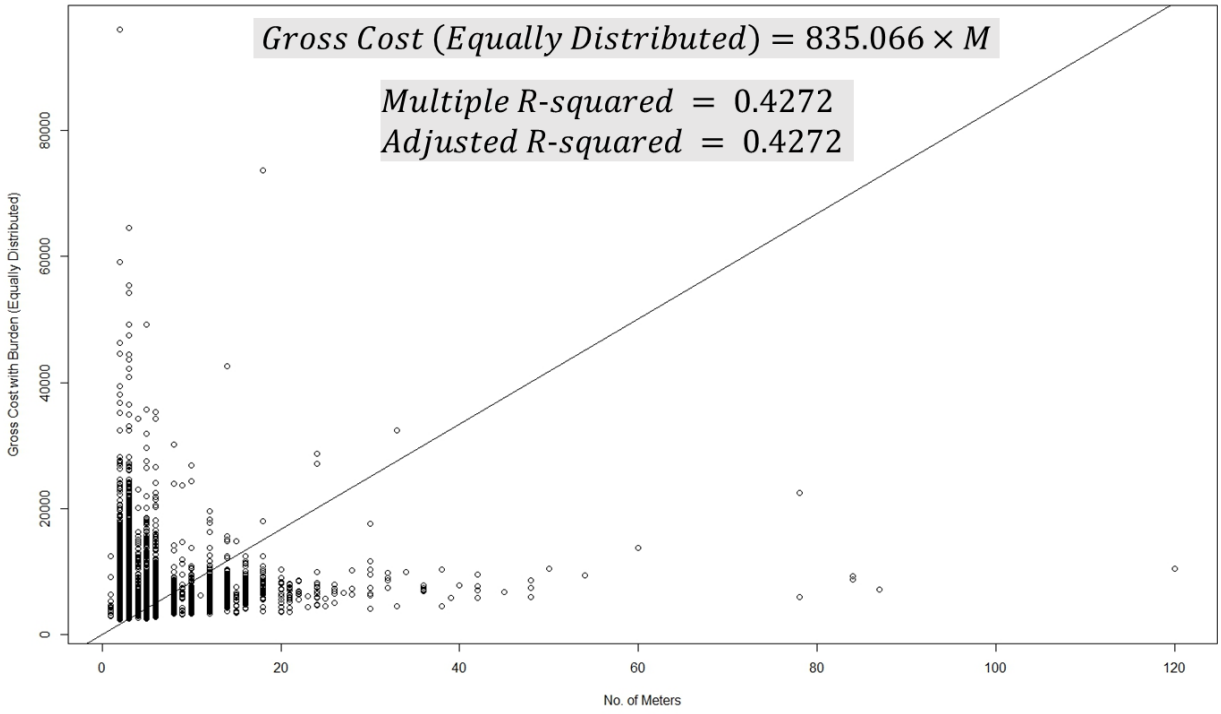

Figure 69. Linear Regression without Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.2.1.1.2 Length of Conductor Cable

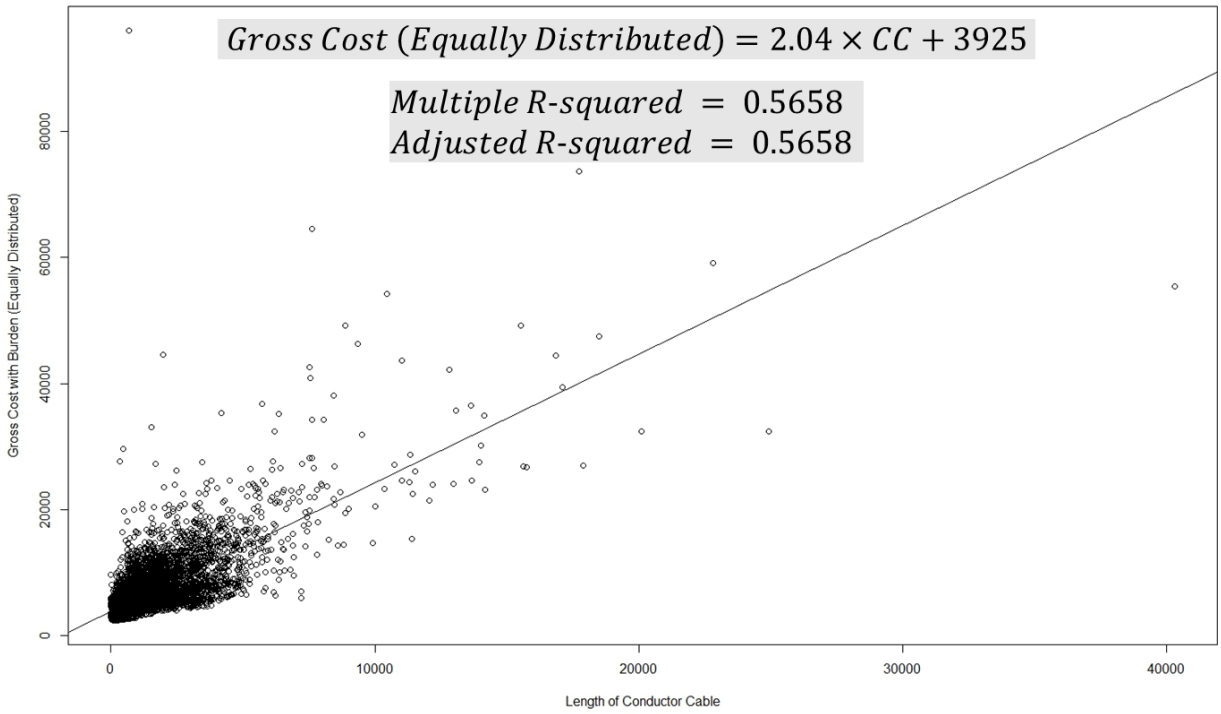

Figure 70. Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

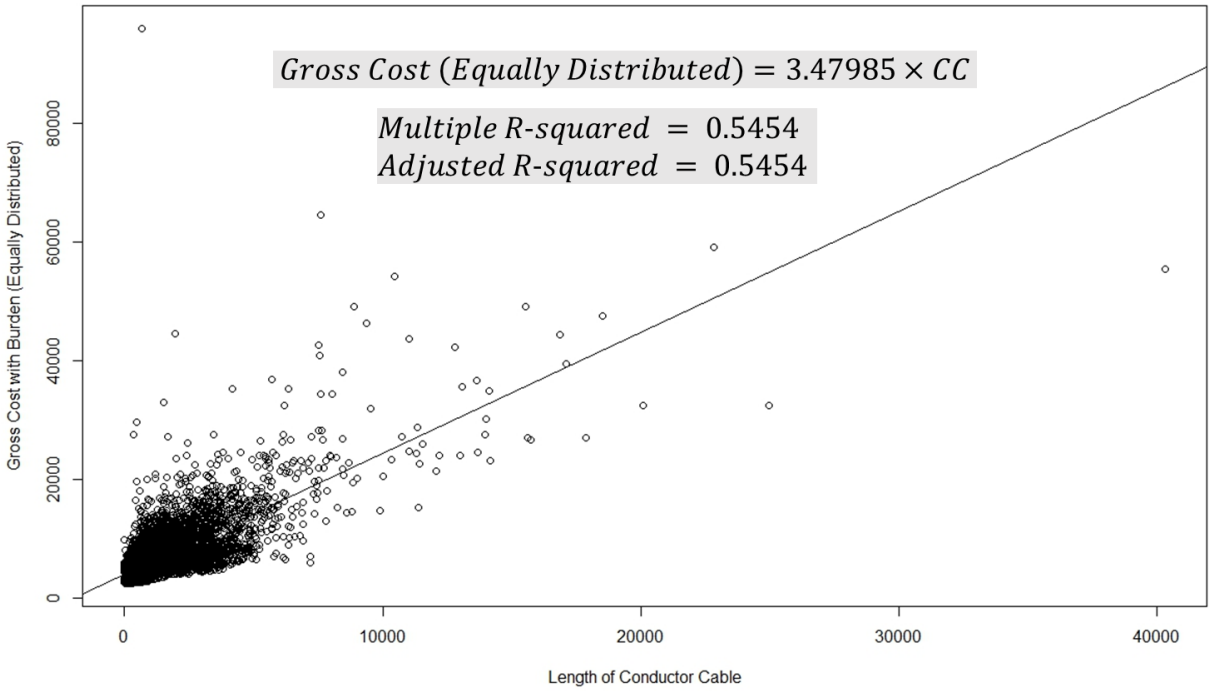

Figure 71. Linear Regression without Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.2.1.2 Multiple Linear Regression

```
Call:
lm(formula = TCE ~ M + CC + TRM + L + TF + CA, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-19138   -747   -358   1093  44291

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3893.43728   17.27994 225.315  < 2e-16 ***
M             32.53080    4.77552   6.812 9.89e-12 ***
CC             1.26133    0.01674  75.371  < 2e-16 ***
TRM            0.75530    0.02280  33.121  < 2e-16 ***
L             29.57076    0.44785  66.028  < 2e-16 ***
TF            86.40397    1.90474  45.363  < 2e-16 ***
CA           -39.67036    3.22364 -12.306  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1546 on 20925 degrees of freedom
Multiple R-squared:  0.7404,    Adjusted R-squared:  0.7404
F-statistic:  9949 on 6 and 20925 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 32.53080 \times M + 1.26133 \times CC + 0.75530 \times TRM + 29.57076 \times L + 86.4397 \times TF - 39.67036 \times CA + 3893.43728$$

$$Multiple\ R\text{-}squared\ = 0.7404$$
$$Adjusted\ R\text{-}squared\ = 0.7404$$

Figure 72. Multiple Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters, Length of Conductor Cable, Trenching Machine, Labor, Transformers, Cable Conn

```
Call:
lm(formula = TCE ~ M + CC + TRM + L + TF + CA - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-38974    906   1701   2808  61945

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
M   170.89848    8.76581  19.496  < 2e-16 ***
CC    0.16103    0.02963   5.435 5.54e-08 ***
TRM   2.56490    0.03950  64.927  < 2e-16 ***
L    36.26829    0.82711  43.849  < 2e-16 ***
TF   89.20738    3.52549  25.304  < 2e-16 ***
CA  290.08210    5.31641  54.564  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2862 on 20926 degrees of freedom
Multiple R-squared:  0.7687,    Adjusted R-squared:  0.7686
F-statistic: 1.159e+04 on 6 and 20926 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 170.89848 \times M + 0.16103 \times CC + 2.56490 \times TRM + 36.26829 \times L + 89.20738 \times TF - 290.08210 \times CA$$

$$Multiple\ R\text{-}squared\ = 0.7687$$
$$Adjusted\ R\text{-}squared\ = 0.7686$$

Figure 73. Multiple Linear Regression without Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters, Length of Conductor Cable, Trenching Machine, Labor, Transformers, Cable Conn

16.2.2 Gross Cost with Burden (Meters)

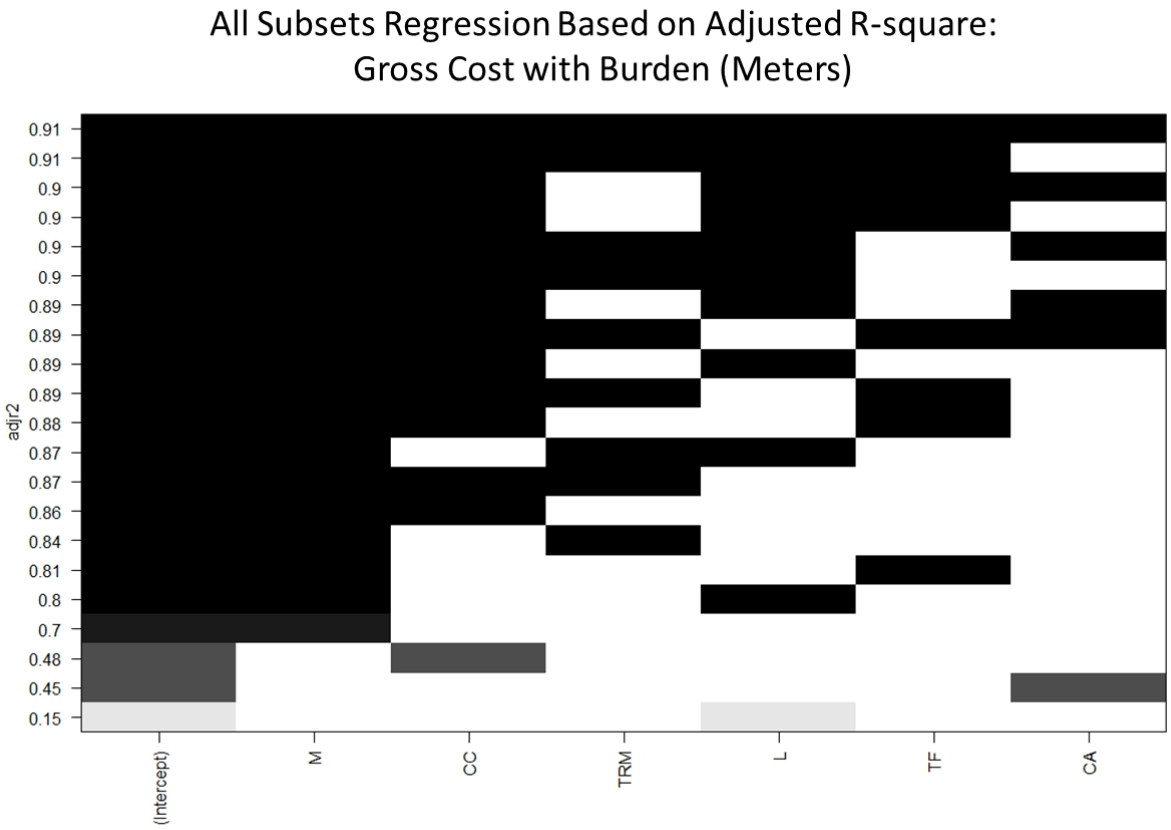

Figure 74. All Subsets Regression Based on Adjusted R-square: Gross Cost with Burden (Meters)

16.2.2.1 Simple Linear Regression

16.2.2.1.1 Meters

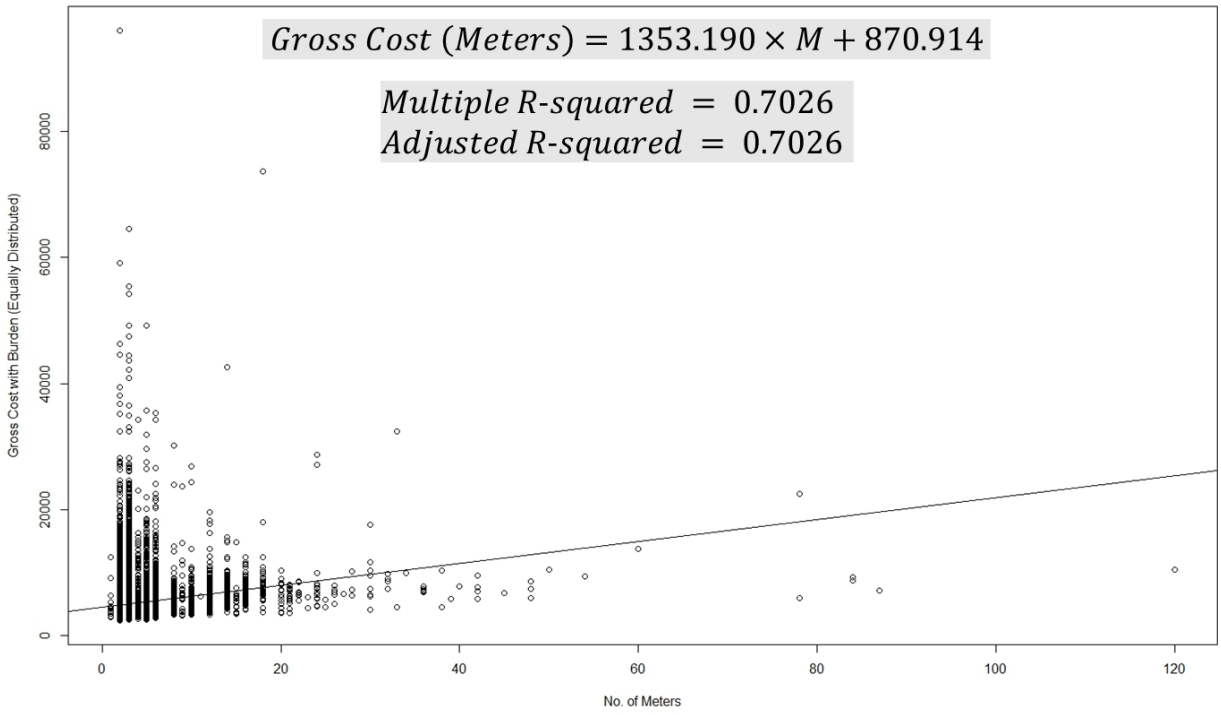

Figure 75. Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

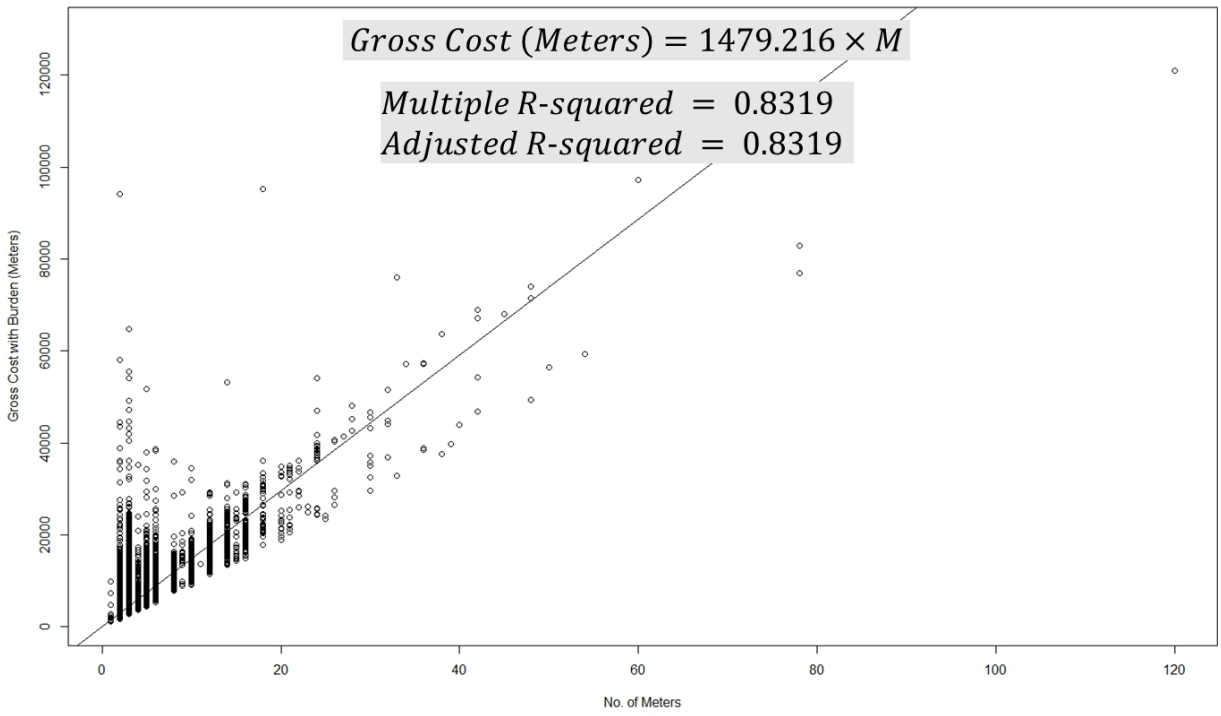

Figure 76. Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.2.1.2 Length of Conductor Cable

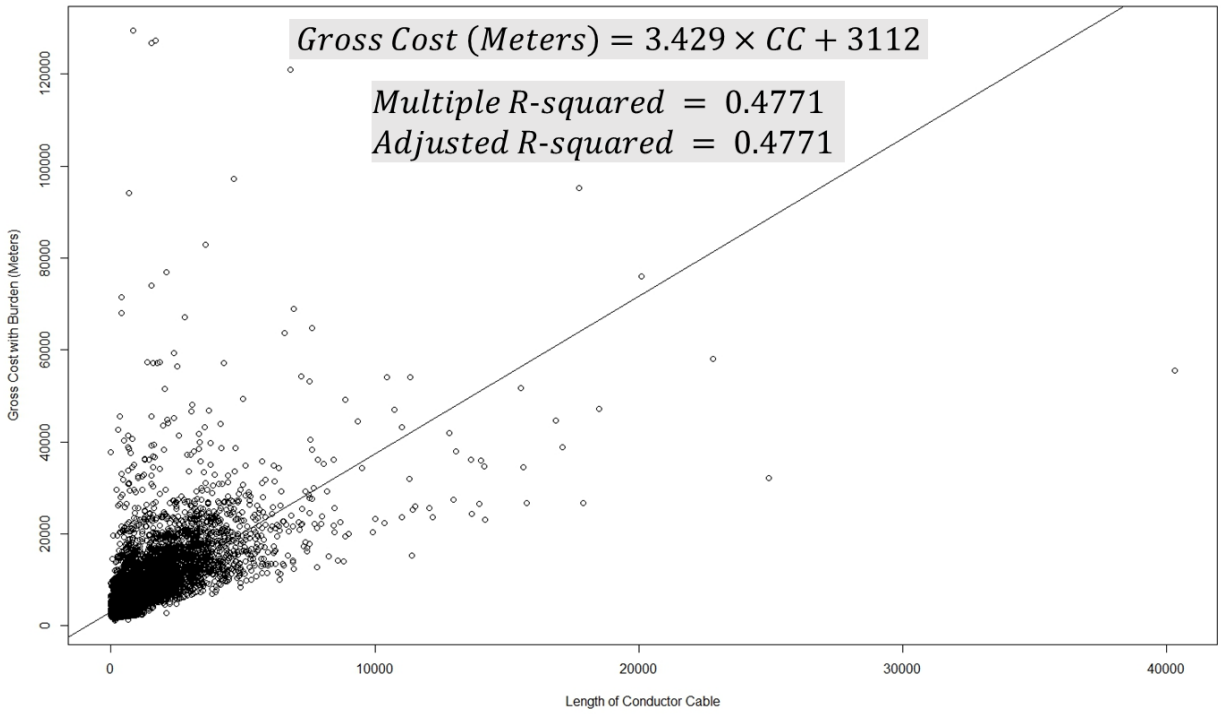

Figure 77. Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

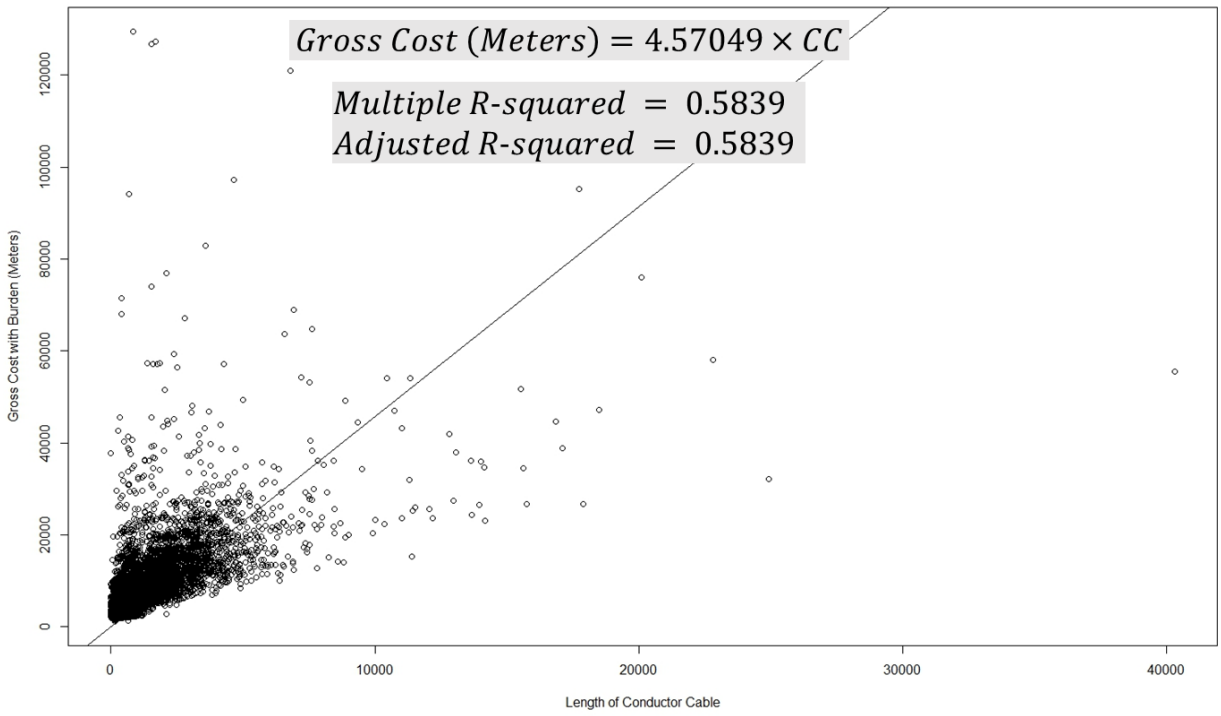

Figure 78. Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.2.2.2 Multiple Linear Regression

```
Call:
lm(formula = TCM ~ M + CC + TRM + L + TF + CA, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-30548   -530   -136    638  47599

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  258.32909   18.92430   13.65   <2e-16 ***
M           1238.74470    5.22995  236.86   <2e-16 ***
CC             1.28714    0.01833   70.23   <2e-16 ***
TRM            0.71403    0.02497   28.59   <2e-16 ***
L             28.89842    0.49047   58.92   <2e-16 ***
TF            85.53791    2.08600   41.01   <2e-16 ***
CA           -63.94786    3.53040  -18.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1693 on 20925 degrees of freedom
Multiple R-squared:  0.9071,    Adjusted R-squared:  0.9071
F-statistic: 3.405e+04 on 6 and 20925 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1238.74470 \times M + 1.28714 \times CC + 0.71403 \times TRM + 28.89842 \times L + 85.53791 \times TF - 63.94786 \times CA + 258.32909$$

$$Multiple\ R\text{-}squared\ = 0.9071$$
$$Adjusted\ R\text{-}squared\ = 0.9071$$

Figure 79. Multiple Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters, Length of Conductor Cable, Trenching Machine, Labor, Transformers, Cable Conn

```
Call:
lm(formula = TCM ~ M + CC + TRM + L + TF + CA - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-32997   -391     -4    772  48770

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
M   1247.92538    5.20945  239.55   <2e-16 ***
CC     1.21414    0.01761   68.96   <2e-16 ***
TRM    0.83410    0.02348   35.53   <2e-16 ***
L     29.34280    0.49155   59.70   <2e-16 ***
TF    85.72392    2.09517   40.91   <2e-16 ***
CA   -42.06882    3.15950  -13.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1701 on 20926 degrees of freedom
Multiple R-squared:  0.9493,    Adjusted R-squared:  0.9493
F-statistic: 6.53e+04 on 6 and 20926 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1247.92538 \times M + 1.21414 \times CC + 0.83410 \times TRM + 29.34280 \times L + 85.72392 \times TF - 42.06882 \times CA$$

$$Multiple\ R\text{-}squared\ = 0.9493$$
$$Adjusted\ R\text{-}squared\ = 0.9493$$

Figure 80. Multiple Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters, Length of Conductor Cable, Trenching Machine, Labor, Transformers, Cable Conn

16.2.3 Gross Cost with Burden (Gross Cost)

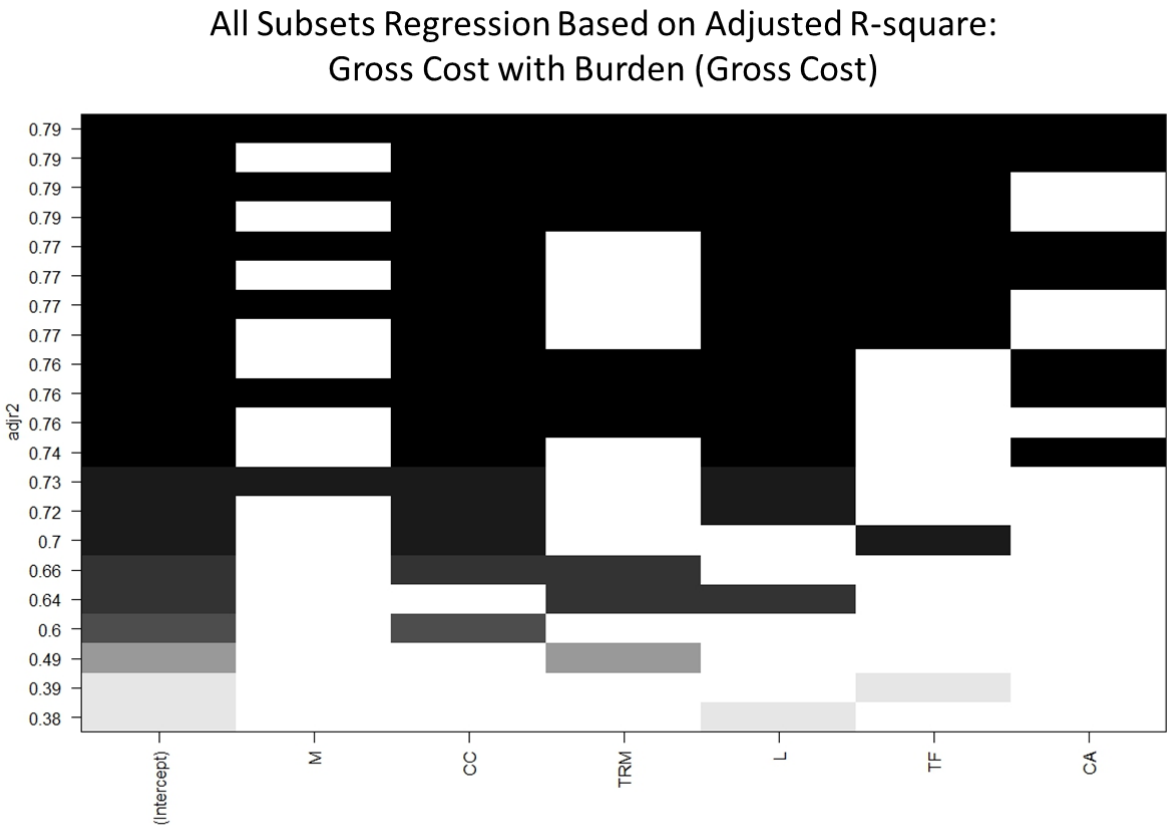

Figure 81. All Subsets Regression Based on Adjusted R-square: Gross Cost with Burden (Gross Cost)

16.2.3.1 Linear Regression

16.2.3.1.1 Meters

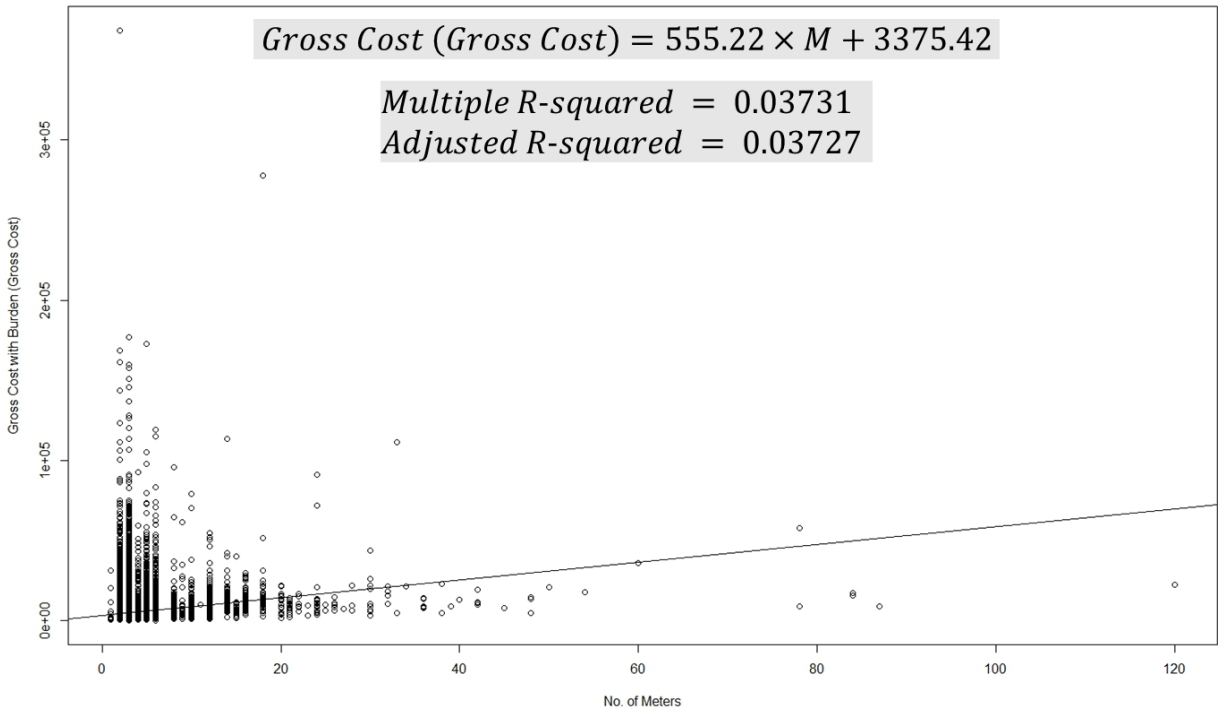

Figure 82. Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

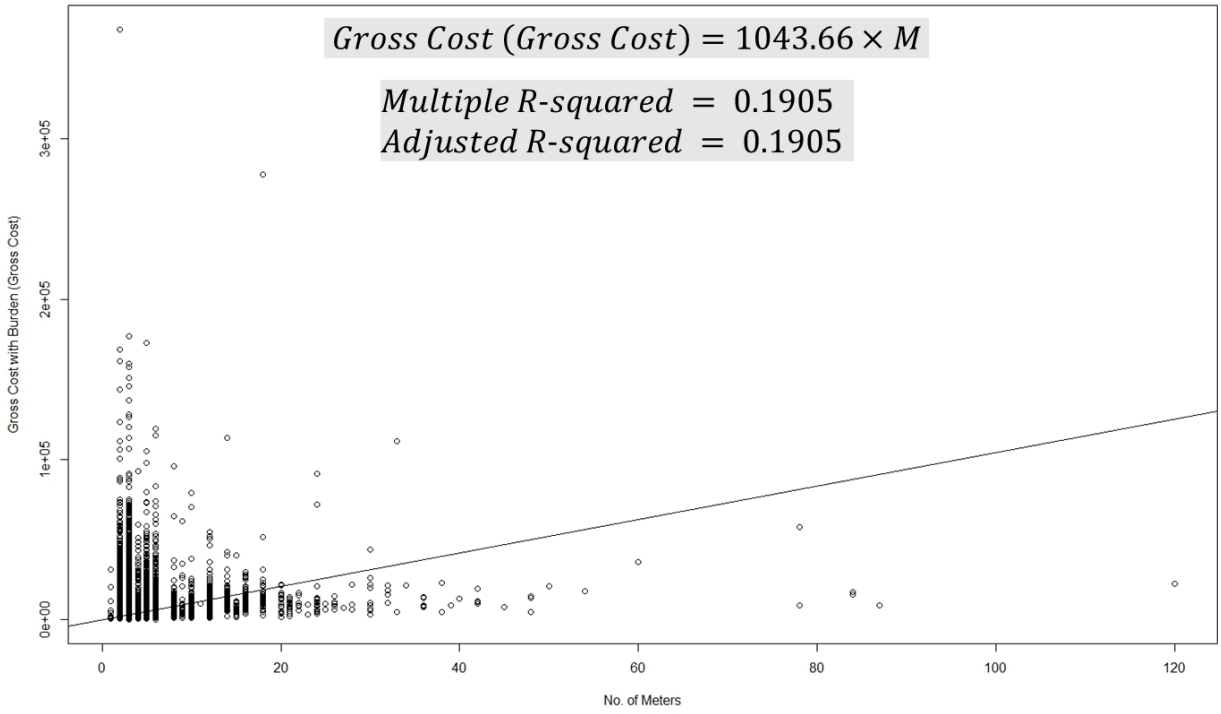

Figure 83. Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.2.3.1.2 Length of Conductor Cable

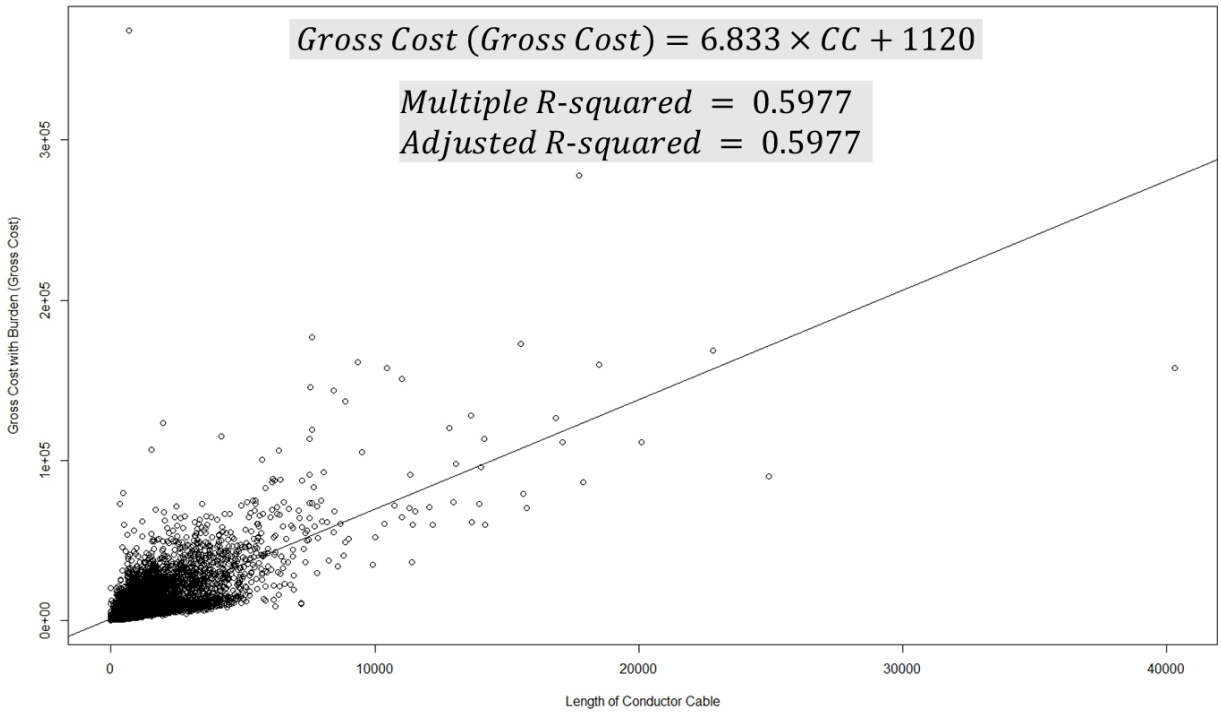

Figure 84. Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

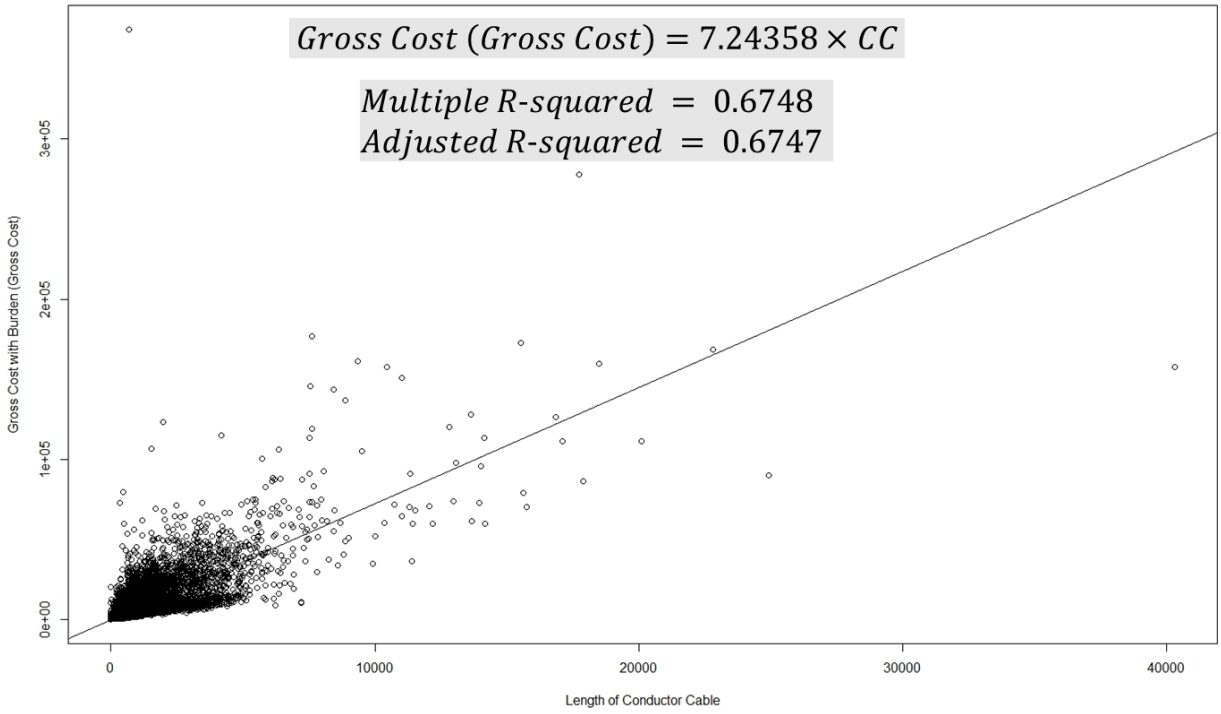

Figure 85. Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.3.2 Multiple Linear Regression

```
Call:
lm(formula = TCG ~ M + CC + TRM + L + TF + CA, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-72683   -552    -31    491 199889

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 965.71108   50.74153  19.032  < 2e-16 ***
M            48.39347   14.02303   3.451  0.00056 ***
CC            4.12320    0.04914  83.905  < 2e-16 ***
TRM           2.65314    0.06696  39.621  < 2e-16 ***
L           101.39725    1.31508  77.103  < 2e-16 ***
TF          291.51556    5.59316  52.120  < 2e-16 ***
CA          -93.12723    9.46602  -9.838  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4541 on 20925 degrees of freedom
Multiple R-squared:  0.7893,    Adjusted R-squared:  0.7892
F-statistic: 1.306e+04 on 6 and 20925 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 48.39347 \times M + 4.12320 \times CC + 2.65314 \times TRM + 101.39725 \times L + 291.51556 \times TF - 93.12723 \times CA + 956.71108$$

$$Multiple\ R\text{-}squared\ = 0.7893$$
$$Adjusted\ R\text{-}squared\ = 0.7892$$

Figure 86. Multiple Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters, Length of Conductor Cable, Trenching Machine, Labor, Transformers, Cable Conn

```
Call:
lm(formula = TCG ~ M + CC + TRM + L + TF + CA - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-69938    -51    475    960 198407

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
M    82.71358   14.02611   5.897 3.76e-09 ***
CC    3.85029    0.04741  81.219  < 2e-16 ***
TRM   3.10198    0.06321  49.074  < 2e-16 ***
L   103.05848    1.32346  77.870  < 2e-16 ***
TF  292.21091    5.64111  51.800  < 2e-16 ***
CA  -11.33688    8.50675  -1.333    0.183
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4580 on 20926 degrees of freedom
Multiple R-squared:  0.8309,    Adjusted R-squared:  0.8309
F-statistic: 1.714e+04 on 6 and 20926 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 82.71358 \times M + 3.85029 \times CC + 3.10198 \times TRM + 103.05848 \times L + 292.21091 \times TF - 11.33688 \times CA$$

$$Multiple\ R\text{-}squared\ = 0.8309$$
$$Adjusted\ R\text{-}squared\ = 0.8309$$

Figure 87. Multiple Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters, Length of Conductor Cable, Trenching Machine, Labor, Transformers, Cable Conn

16.2.4 Cross Validation (2005-2015)

16.2.4.1 Gross Cost with Burden (Meters)

16.2.4.1.1 Meters without Intercept

16.2.4.1.1.1 2005 Removed

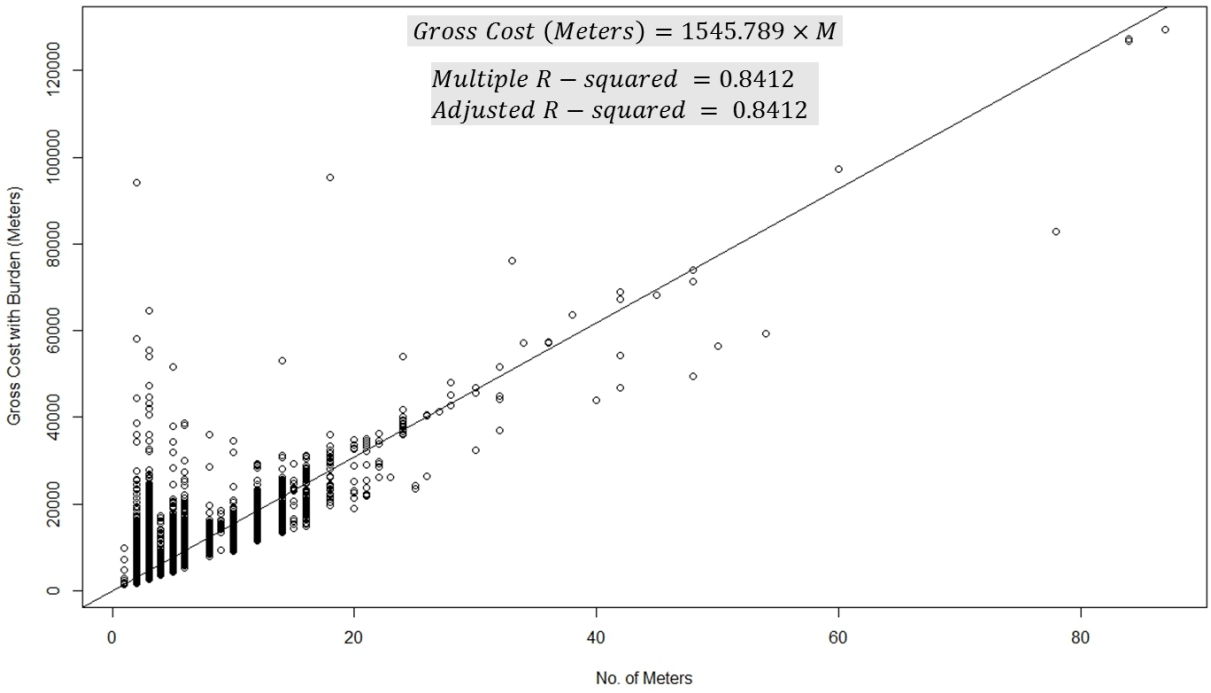

Figure 88. 2005 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.2 2006 Removed

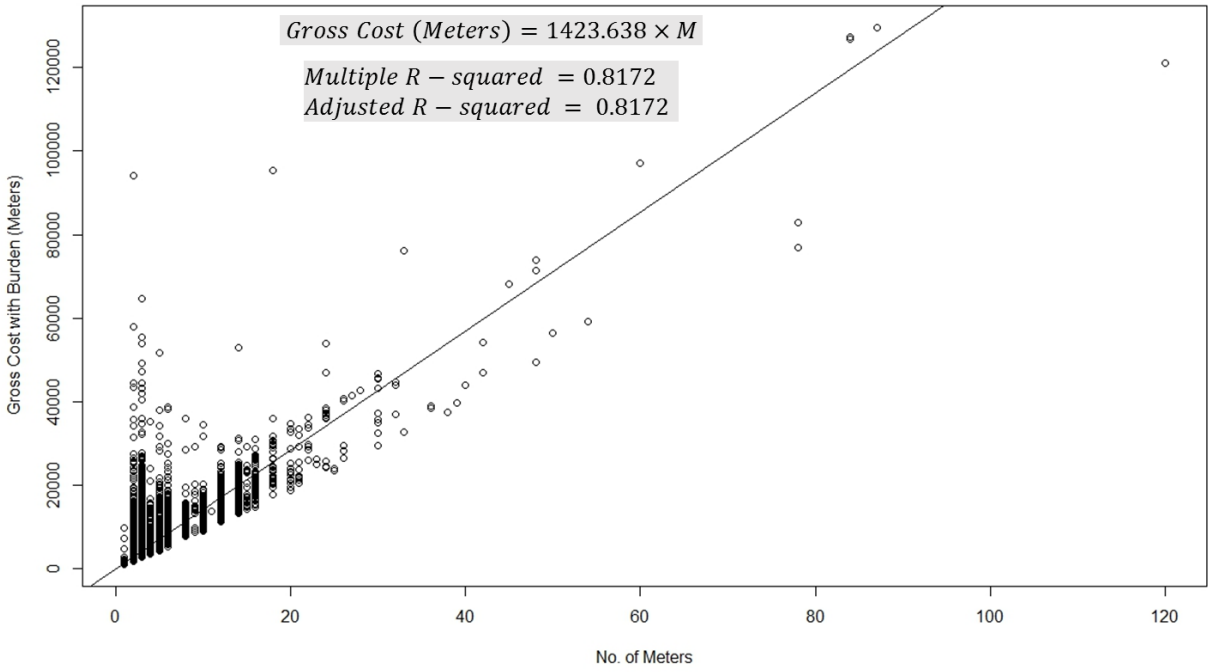

Figure 89. 2006 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.3 2007 Removed

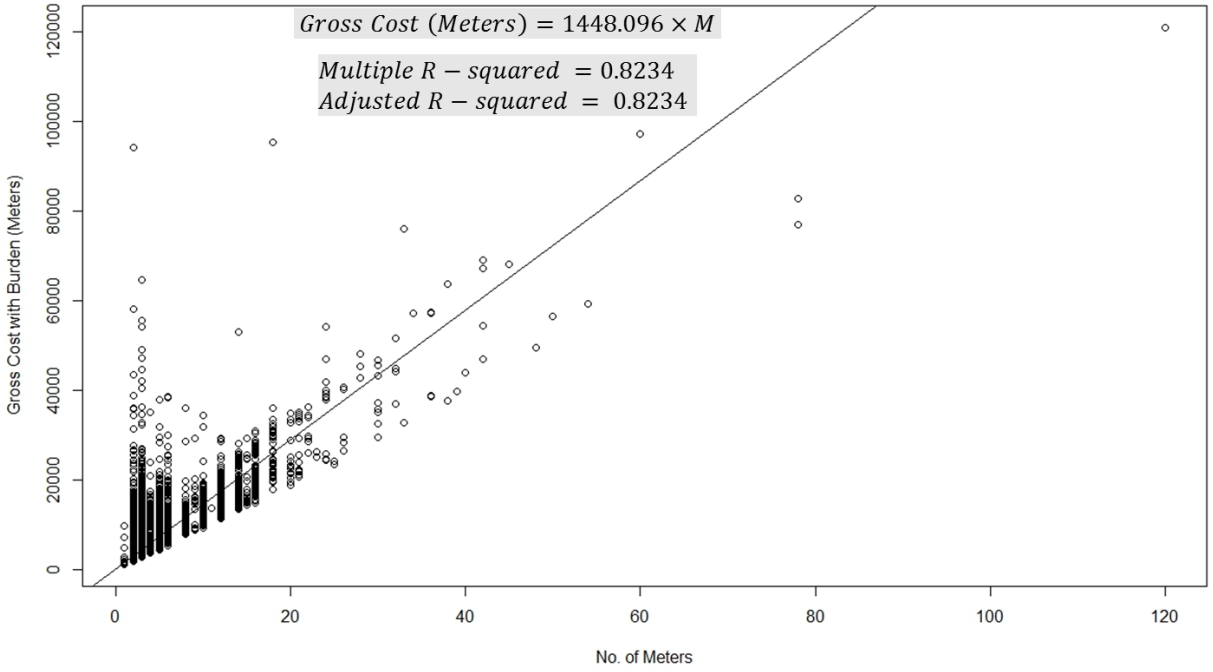

Figure 90. 2007 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.4 2008 Removed

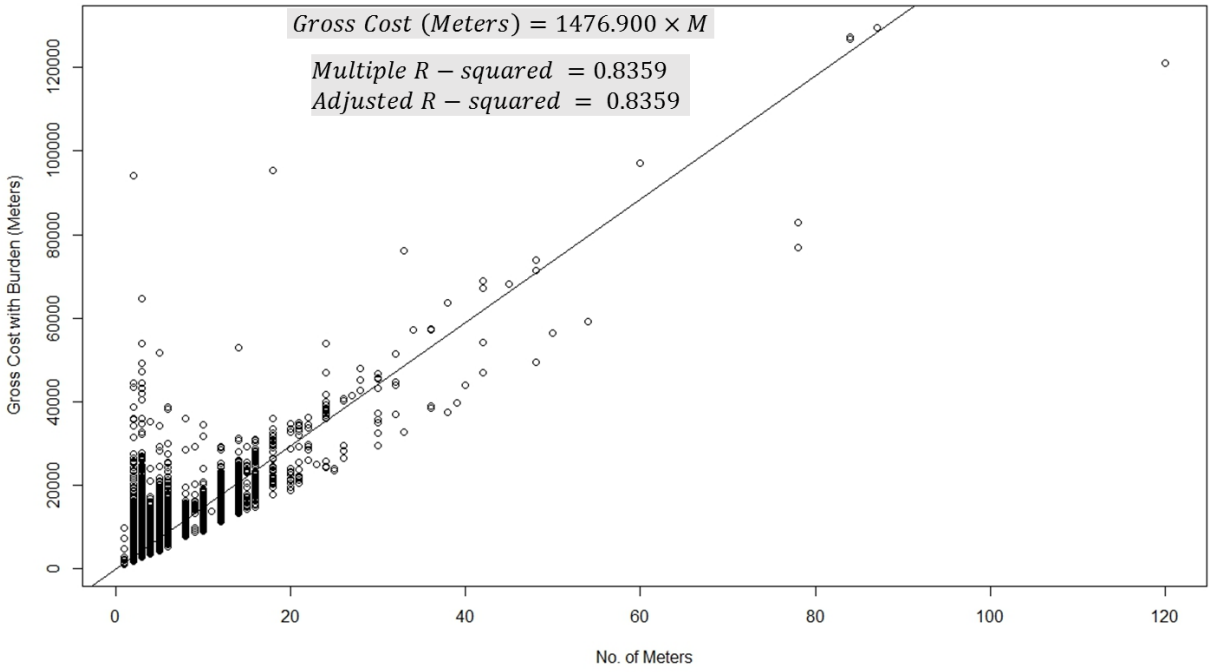

Figure 91. 2008 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.5 2009 Removed

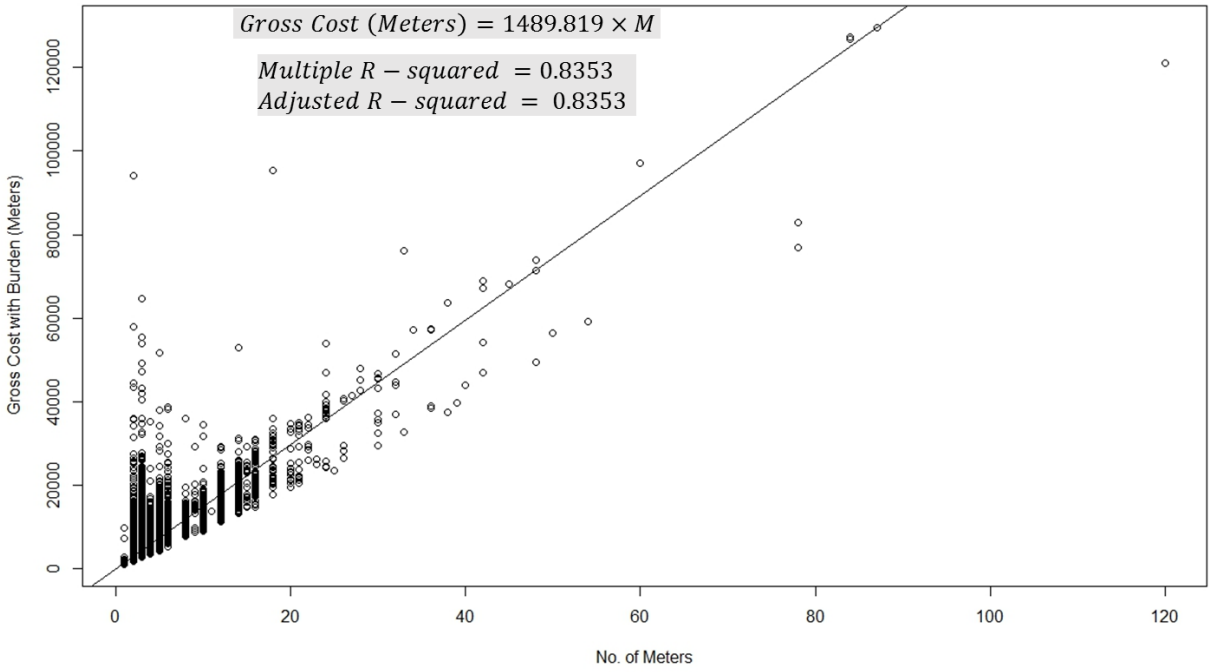

Figure 92. 2009 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.6 2010 Removed

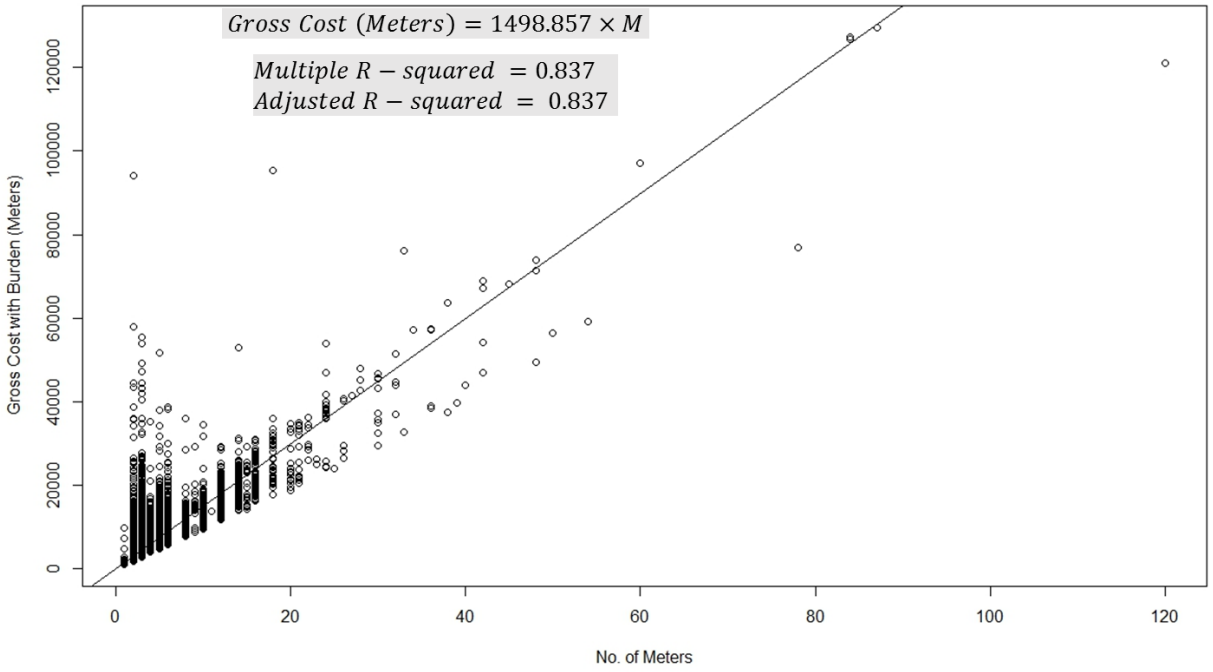

Figure 93. 2010 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.7 2011 Removed

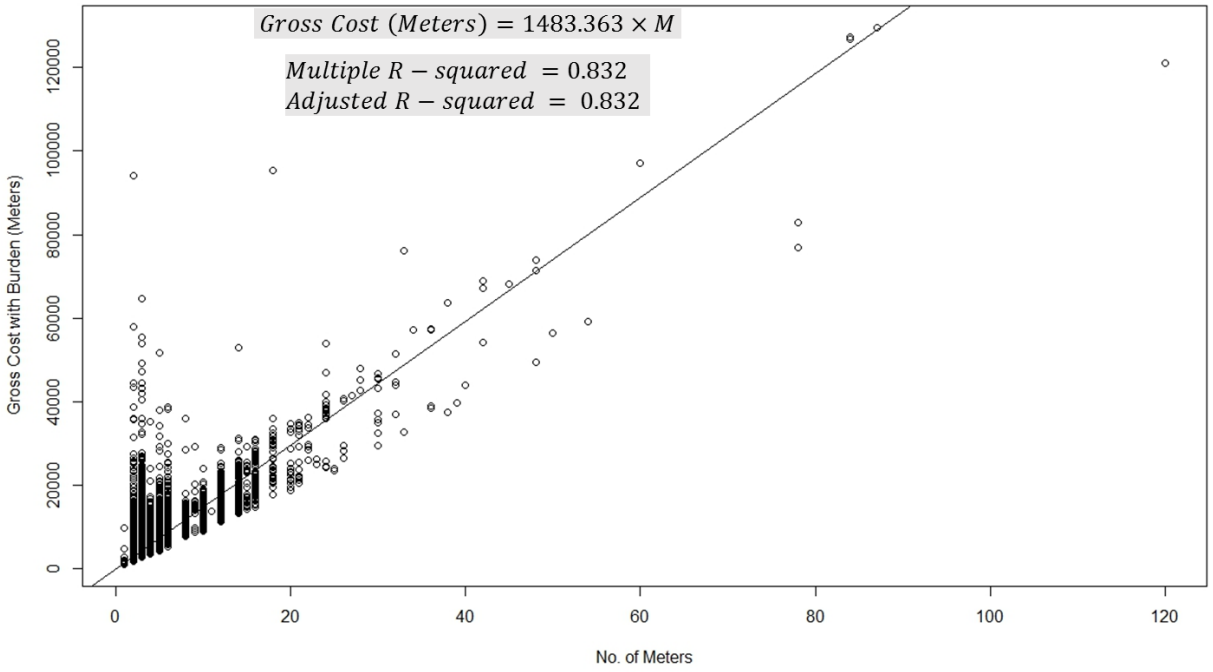

Figure 94. 2011 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.8 2012 Removed

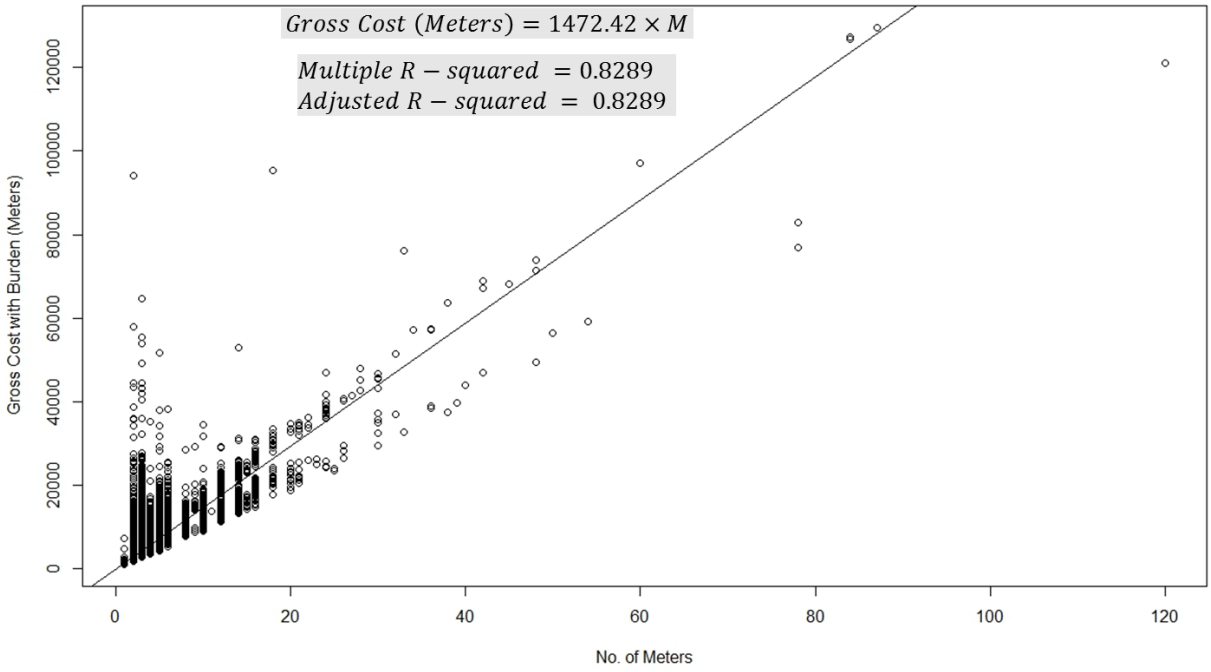

Figure 95. 2012 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.9 2013 Removed

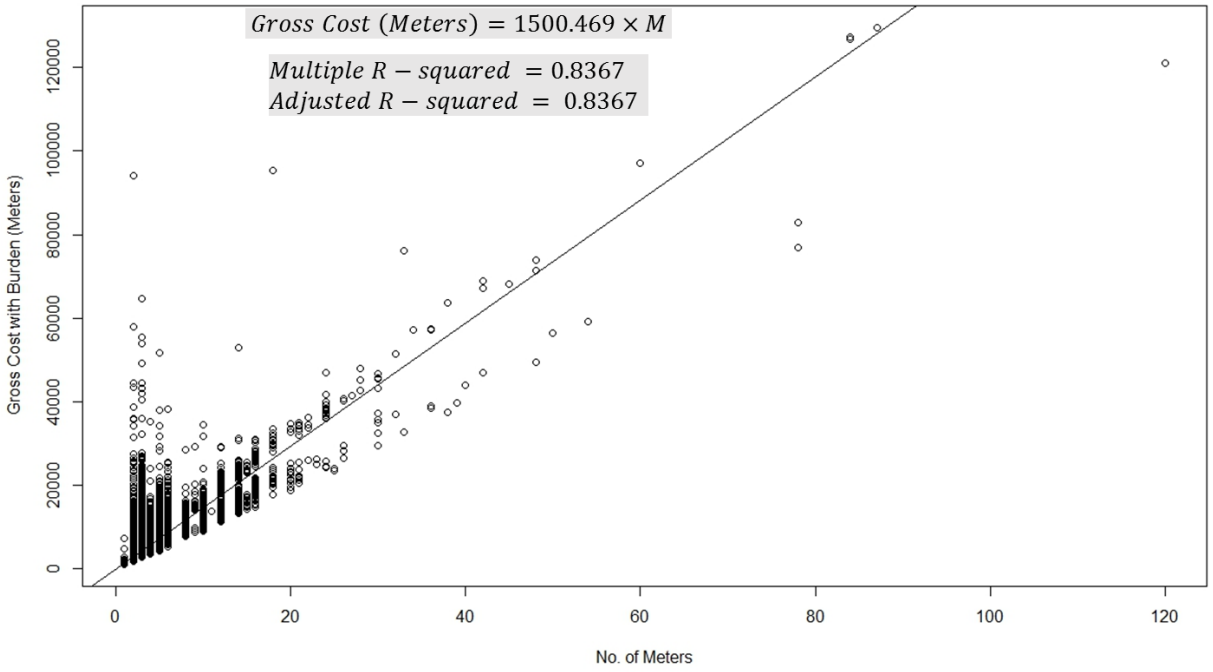

Figure 96. 2013 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.10 2014 Removed

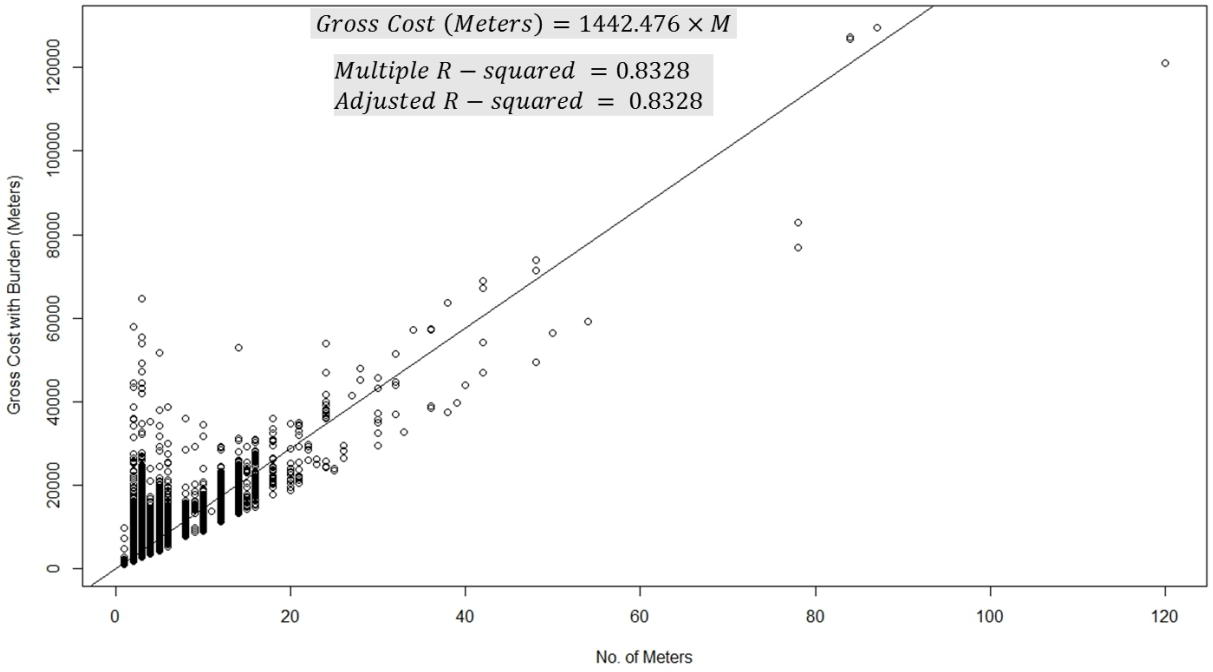

Figure 97. 2014 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.1.1.11 2015 Removed

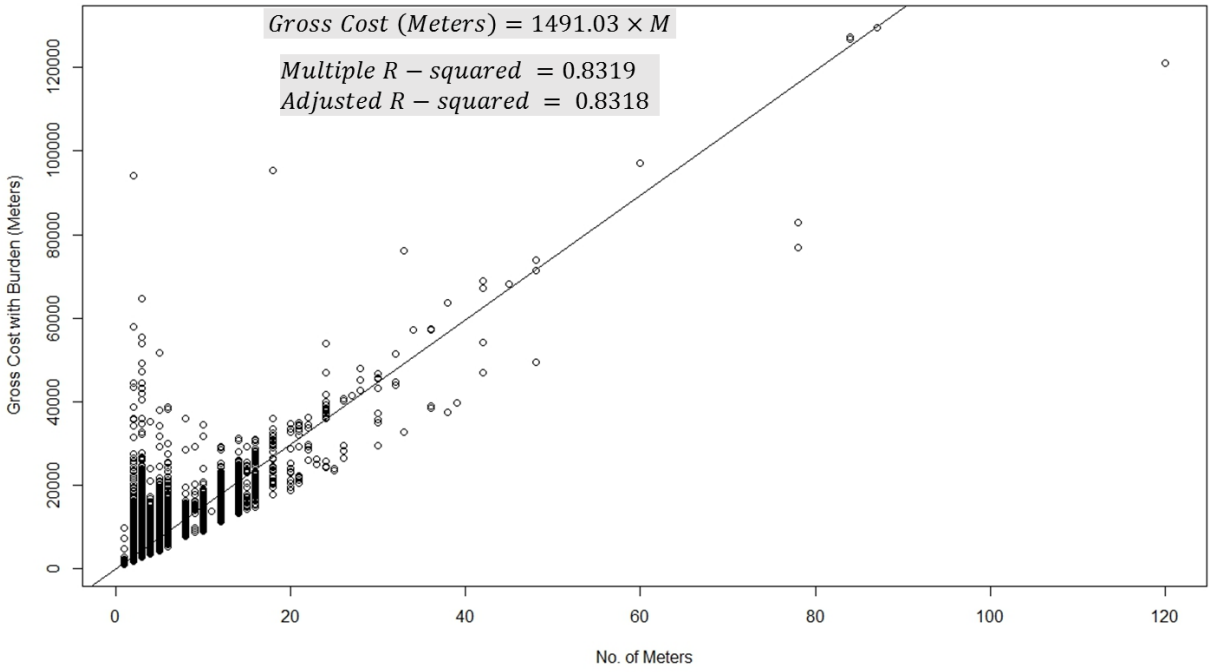

Figure 98. 2015 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.2.4.2 Gross Cost with Burden (Gross Cost)

16.2.4.2.1 Length of Conductor Cable without Intercept

16.2.4.2.1.1 2005 Removed

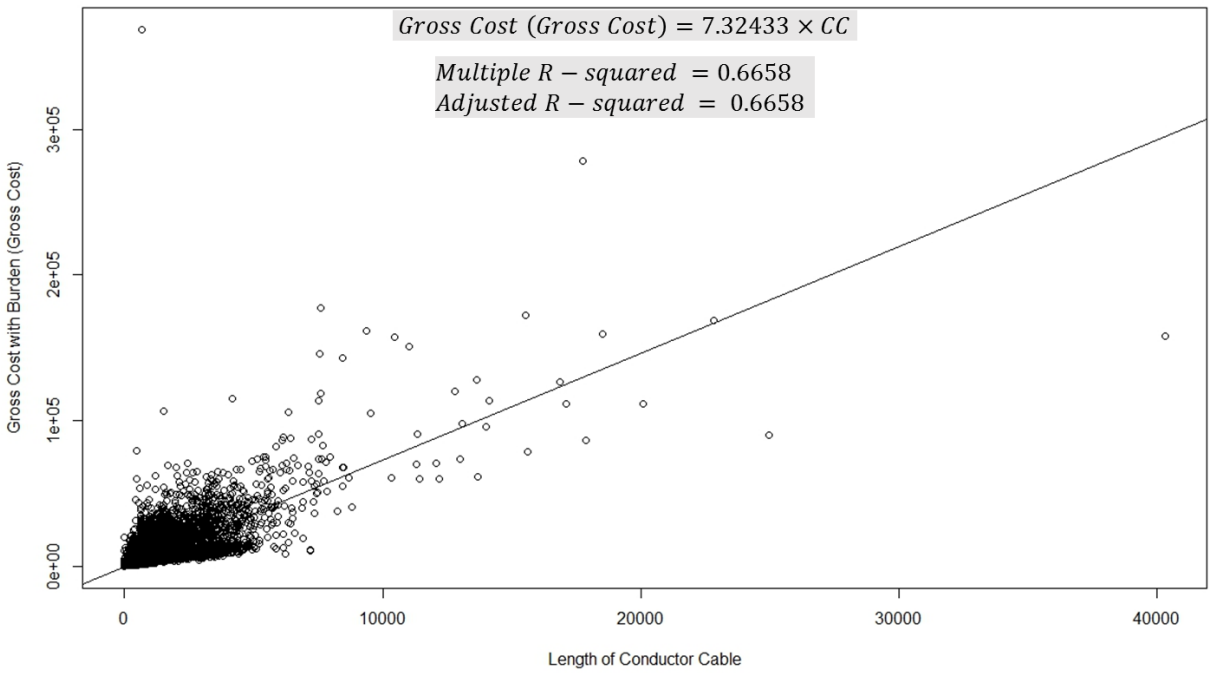

Figure 99. 2005 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.2 2006 Removed

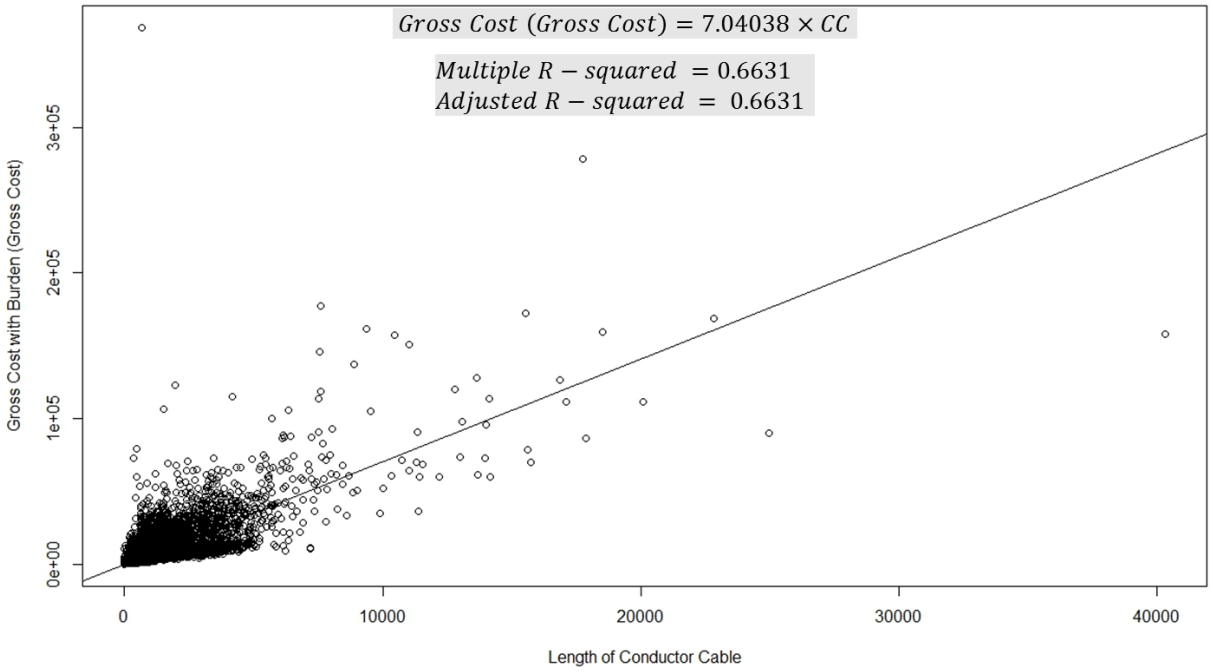

Figure 100. 2006 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.3 2007 Removed

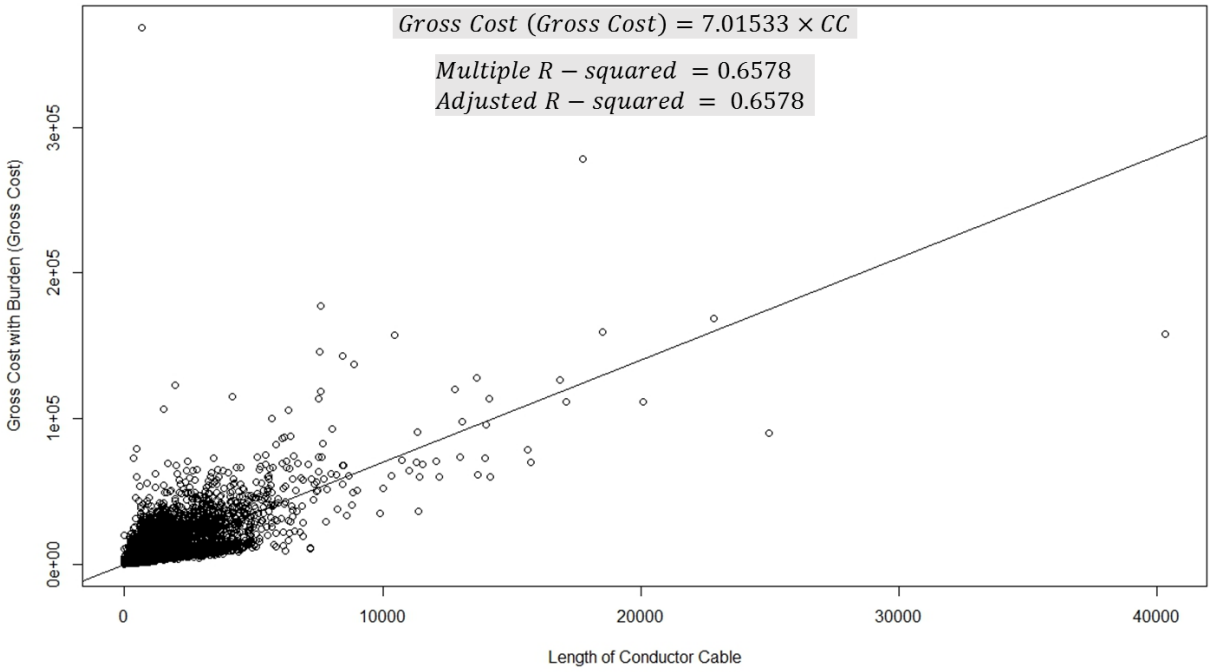

Figure 101. 2007 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.4 2008 Removed

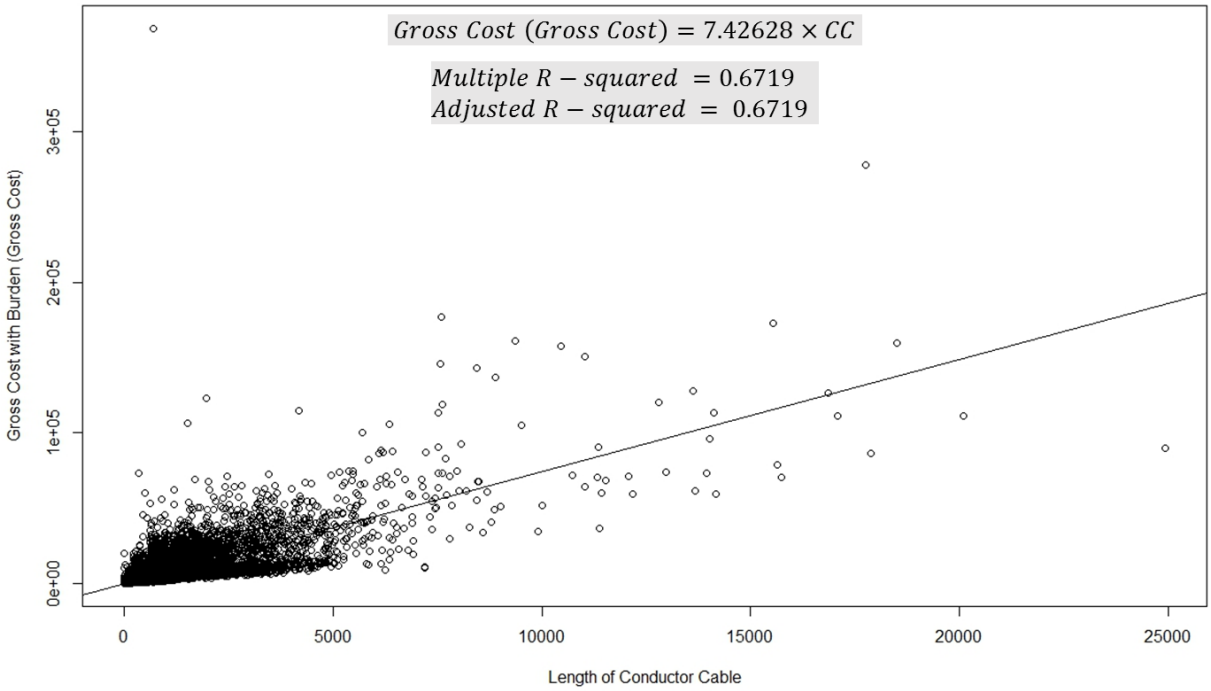

Figure 102. 2008 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.5 2009 Removed

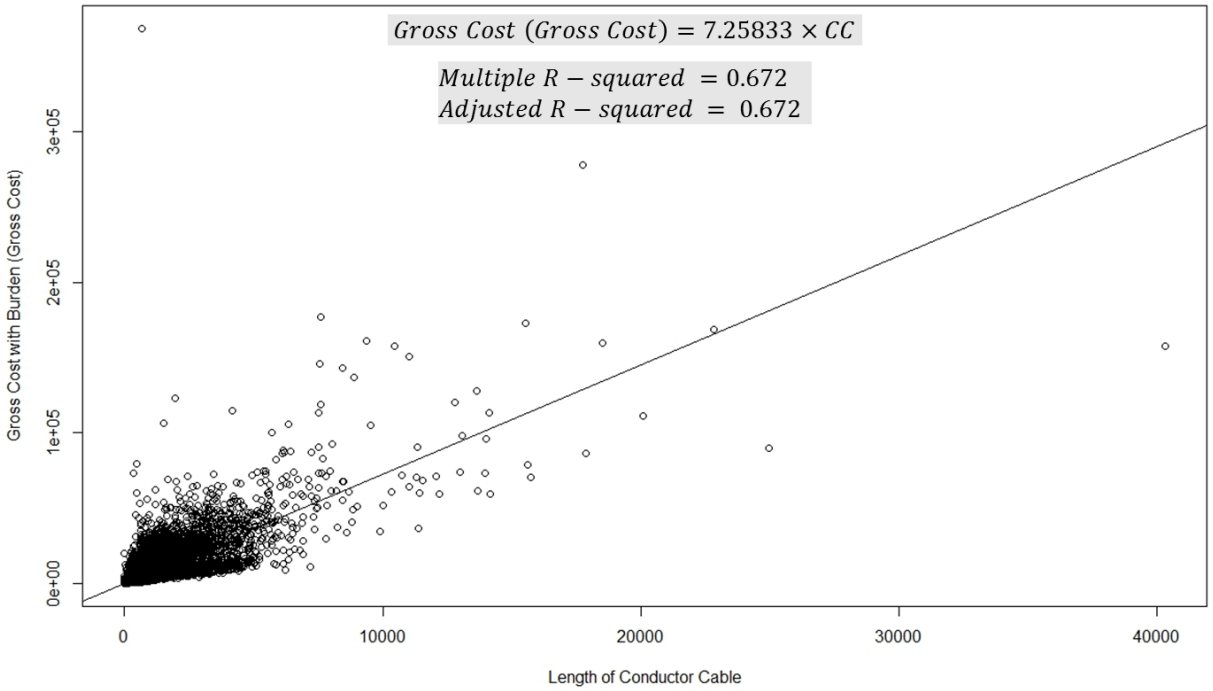

Figure 103. 2009 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.6 2010 Removed

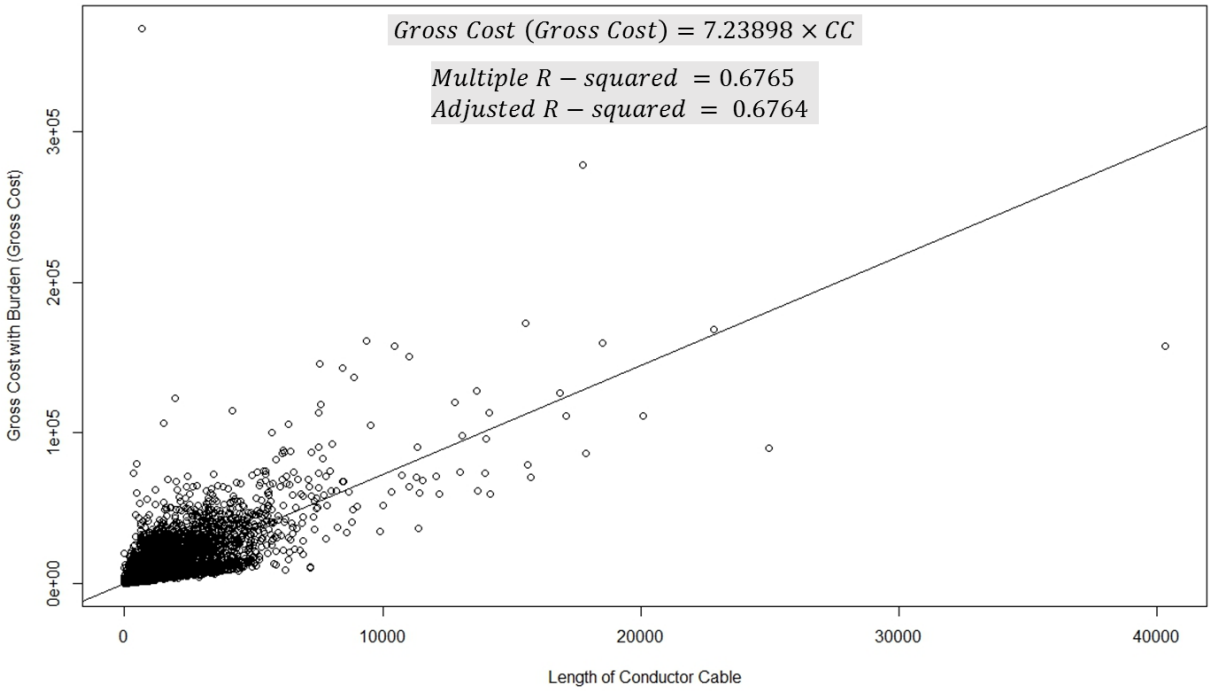

Figure 104. 2010 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.7 2011 Removed

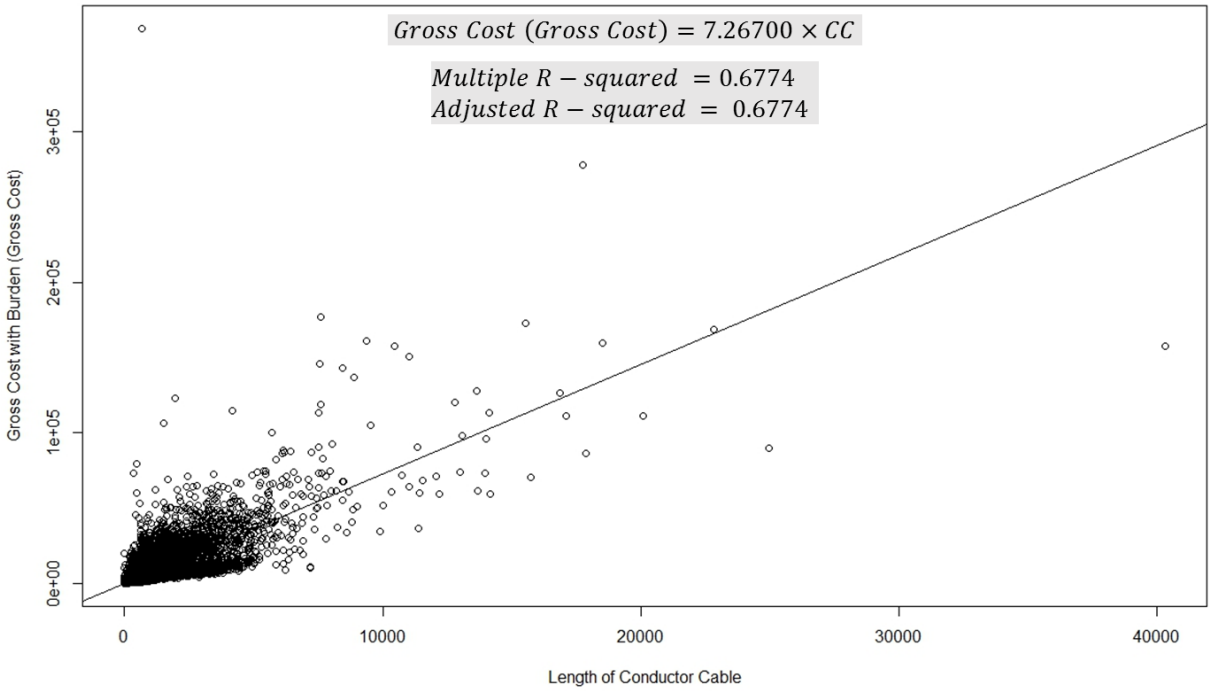

Figure 105. 2011 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.8 2012 Removed

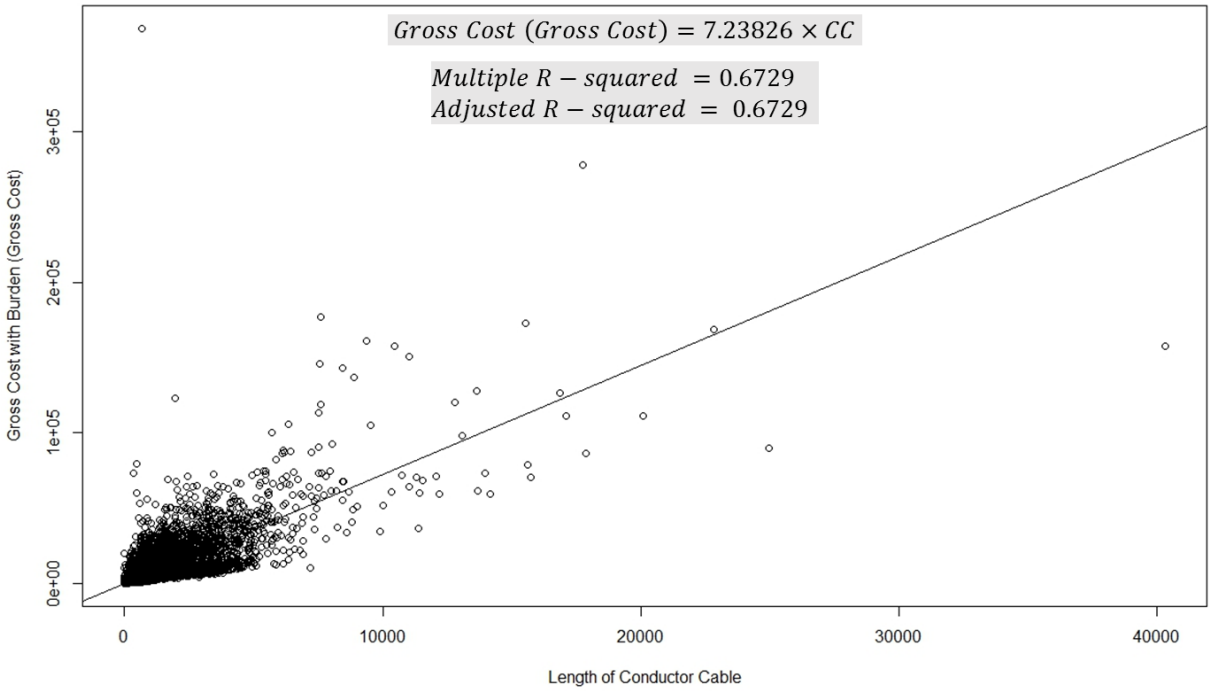

Figure 106. 2012 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.9 2013 Removed

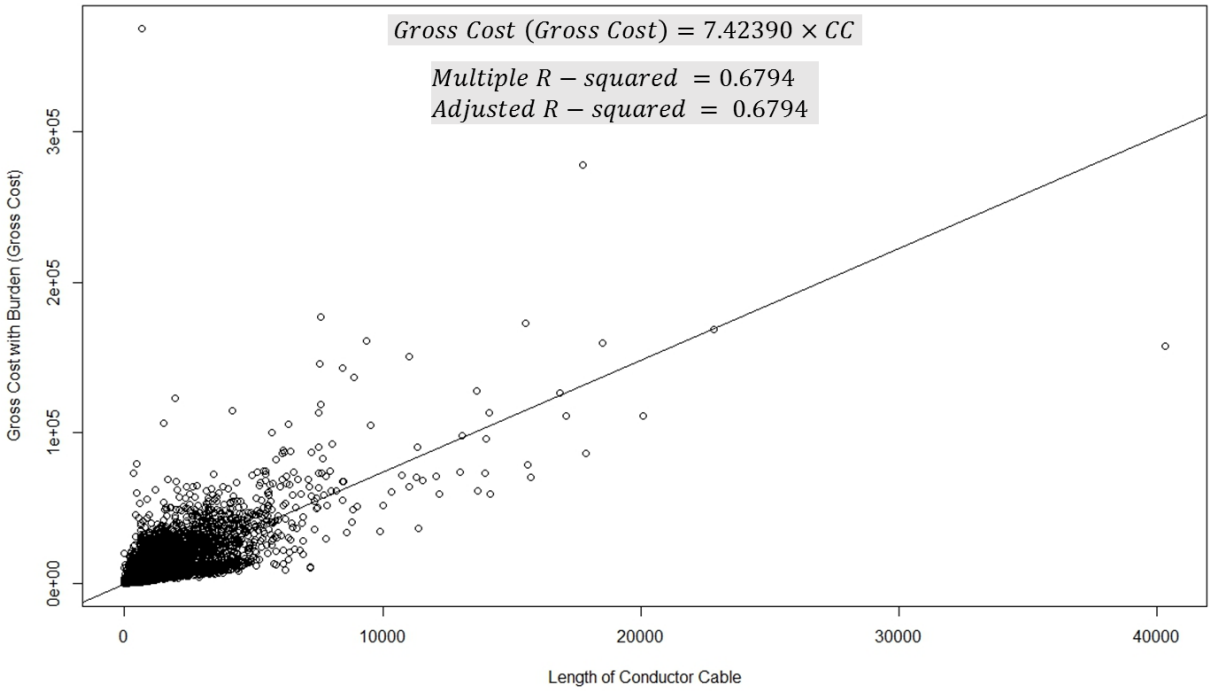

Figure 107. 2013 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.10 2014 Removed

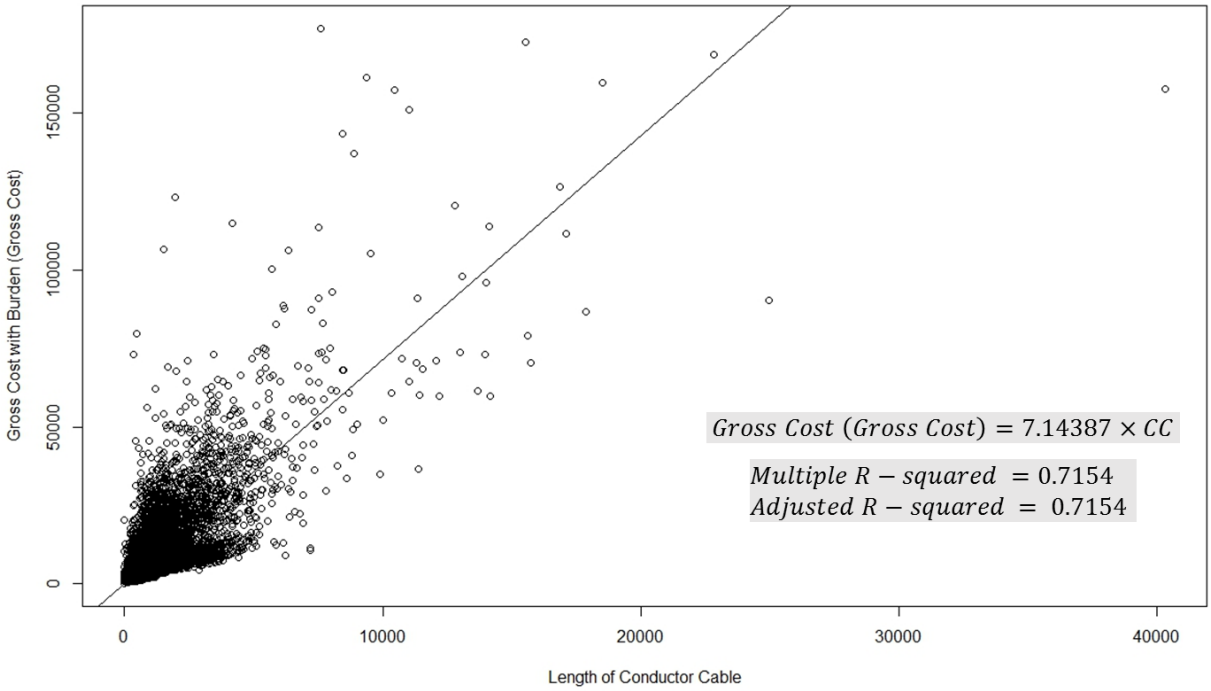

Figure 108. 2014 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.4.2.1.11 2015 Removed

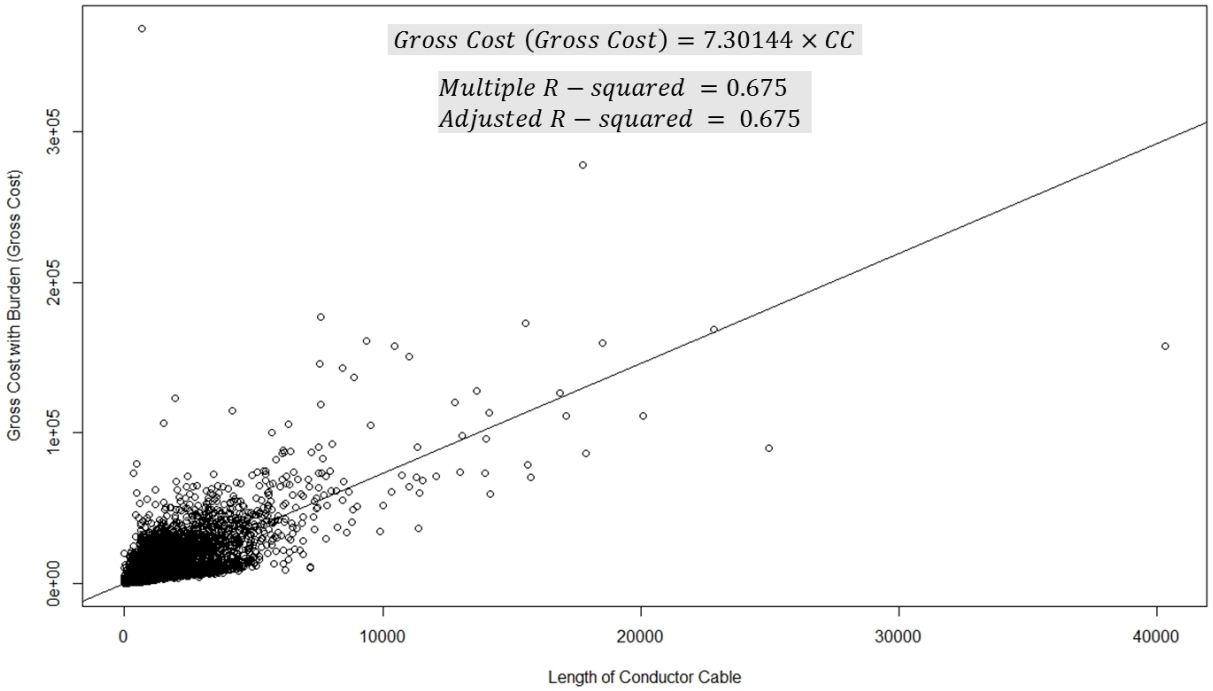

Figure 109. 2015 Removed, Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.2.5 Other Models: 2005 - 2015 Initial Validation Results

		No. of Meters					
	Actual	Predicted Gross Cost with Burden (Equally Distributed) Predicted from No. of Meters w/ Intercept	Predicted Gross Cost with Burden (Equally Distributed) Predicted from No. of Meters w/o Intercept	Predicted Gross Cost with Burden (Meters) Predicted from No. of Meters w/ Intercept	Predicted Gross Cost with Burden (Meters) Predicted from No. of Meters w/o Intercept	Predicted Gross Cost with Burden (Gross Cost) Predicted from No. of Meters w/ Intercept	Predicted Gross Cost with Burden (Gross Cost) Predicted from No. of Meters w/o Intercept
2005	$18.6 M	$21.3 M	$10.9 M	$21.3 M	$19.3 M	$21.3 M	$13.6 M
2006	$19.2 M	$15.7 M	$8.0 M	$15.7 M	$14.2 M	$15.7 M	$10.0 M
2007	$12.6 M	$10.4 M	$5.3 M	$10.4 M	$9.5 M	$10.4 M	$6.7 M
2008	$6.5 M	$6.5 M	$3.3 M	$6.5 M	$5.9 M	$6.5 M	$4.2 M
2009	$5.0 M	$6.2 M	$3.2 M	$6.2 M	$5.6 M	$6.2 M	$4.0 M
2010	$5.2 M	$6.8 M	$3.5 M	$6.8 M	$6.2 M	$6.8 M	$4.4 M
2011	$6.8 M	$7.4 M	$3.8 M	$7.4 M	$6.7 M	$7.4 M	$4.7 M
2012	$9.1 M	$8.9 M	$4.5 M	$8.9 M	$8.0 M	$8.9 M	$5.7 M
2013	$7.8 M	$9.2 M	$4.7 M	$9.2 M	$8.3 M	$9.2 M	$5.9 M
2014	$11.7 M	$9.7 M	$5.0 M	$9.7 M	$8.8 M	$9.7 M	$6.2 M
	2005	14%	41%	14%	4%	14%	27%
	2006	18%	58%	18%	26%	18%	48%
	2007	17%	58%	17%	25%	17%	47%
	2008	0%	49%	0%	9%	0%	36%
	2009	23%	37%	23%	12%	23%	21%
	2010	32%	32%	32%	20%	32%	16%
	2011	9%	44%	9%	1%	9%	30%
	2012	3%	50%	3%	12%	3%	38%
	2013	18%	40%	18%	7%	18%	25%
	2014	17%	57%	17%	24%	17%	47%
		Percent Error					
	Max % Error	32%	58%	32%	26%	32%	48%
	Avg % Error	15%	47%	15%	14%	15%	33%
	Mean Squared Error	3.40E+12	3.37E+13	3.40E+12	4.59E+12	3.40E+12	2.00E+13

Figure 110. 2005 - 2015 Initial Model Validation Results predicted from No. of Meters

	Length of Conductor Cable					
Actual	Predicted Gross Cost with Burden (Equally Distributed) Predicted from Length of Conductor Cable w/ Intercept	Predicted Gross Cost with Burden (Equally Distributed) Predicted from Length of Conductor Cable w/o Intercept	Predicted Gross Cost with Burden (Meters) Predicted from Length of Conductor Cable w/ Intercept	Predicted Gross Cost with Burden (Meters) Predicted from Length of Conductor Cable w/o Intercept	Predicted Gross Cost with Burden (Gross Cost) Predicted from Length of Conductor Cable w/ Intercept	Predicted Gross Cost with Burden (Gross Cost) Predicted from Length of Conductor Cable w/o Intercept
$18.6 M	$21.3 M	$8.5 M	$21.3 M	$11.1 M	$21.3 M	$17.6 M
$19.2 M	$15.7 M	$6.2 M	$15.7 M	$8.2 M	$15.7 M	$13.0 M
$12.6 M	$10.4 M	$4.1 M	$10.4 M	$5.4 M	$10.4 M	$8.6 M
$6.5 M	$6.5 M	$2.6 M	$6.5 M	$3.4 M	$6.5 M	$5.4 M
$5.0 M	$6.2 M	$2.5 M	$6.2 M	$3.2 M	$6.2 M	$5.1 M
$5.2 M	$6.8 M	$2.7 M	$6.8 M	$3.6 M	$6.8 M	$5.7 M
$6.8 M	$7.4 M	$2.9 M	$7.4 M	$3.9 M	$7.4 M	$6.1 M
$9.1 M	$8.9 M	$3.5 M	$8.9 M	$4.6 M	$8.9 M	$7.3 M
$7.8 M	$9.2 M	$3.7 M	$9.2 M	$4.8 M	$9.2 M	$7.6 M
$11.7 M	$9.7 M	$3.9 M	$9.7 M	$5.1 M	$9.7 M	$8.1 M
2005	14%	55%	14%	40%	14%	5%
2006	18%	67%	18%	57%	18%	32%
2007	17%	67%	17%	57%	17%	31%
2008	0%	60%	0%	48%	0%	17%
2009	23%	51%	23%	36%	23%	2%
2010	32%	48%	32%	31%	32%	9%
2011	9%	57%	9%	43%	9%	10%
2012	3%	61%	3%	49%	3%	19%
2013	18%	53%	18%	39%	18%	3%
2014	17%	67%	17%	56%	17%	31%
	Percent Error					
Max % Error	32%	67%	32%	57%	32%	32%
Avg % Error	15%	59%	15%	46%	15%	16%
Mean Squared Error	3.40E+12	4.94E+13	3.40E+12	3.24E+13	5.89E+11	8.37E+10

Figure 111. 2005 - 2015 Initial Model Validation Results predicted from Length of Conductor Cable

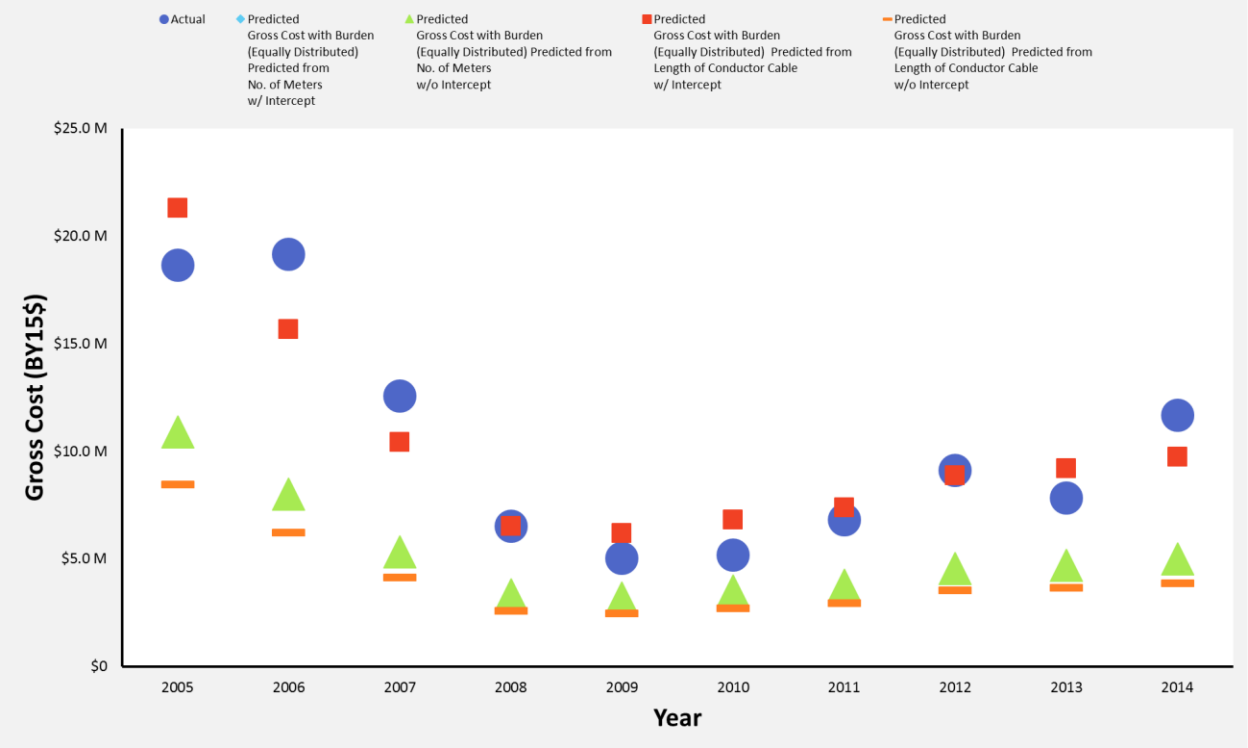

Figure 112. 2005 - 2015 Initial Validation Results for Gross Cost with Burden (Equally Distributed) Models

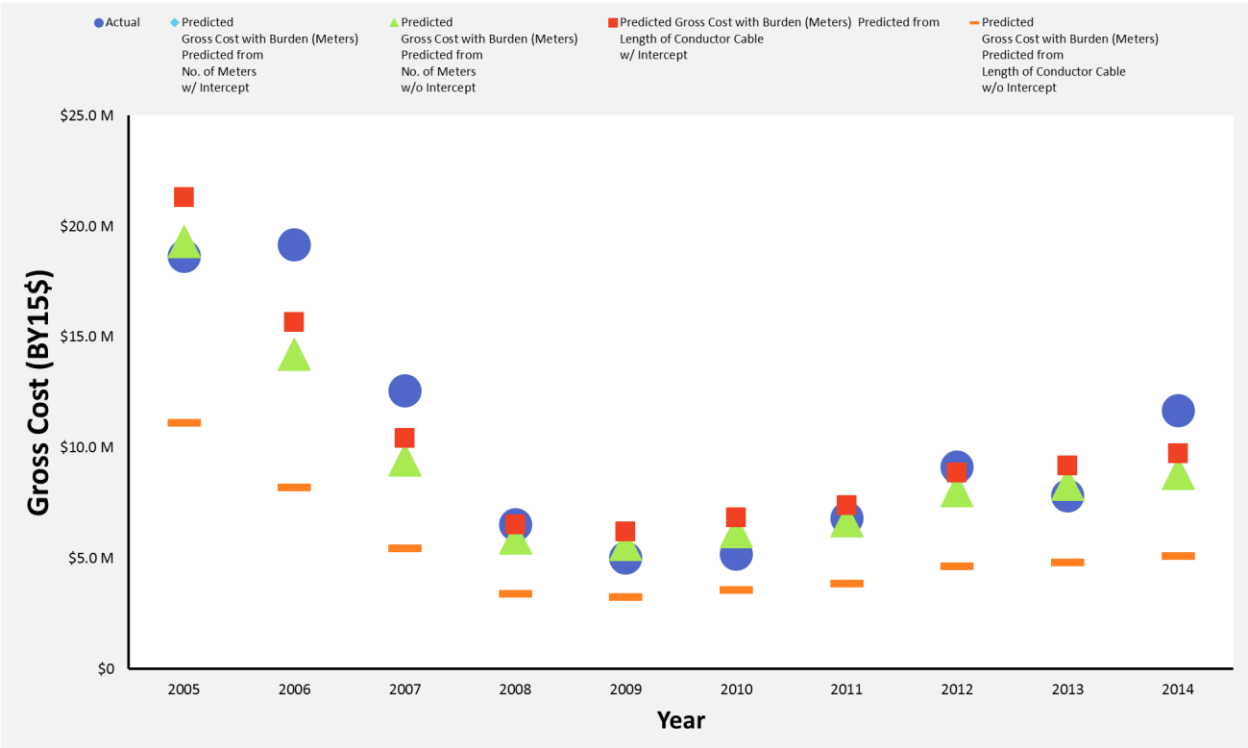

Figure 113. 2005 - 2015 Initial Validation Results for Gross Cost with Burden (Meters) Models

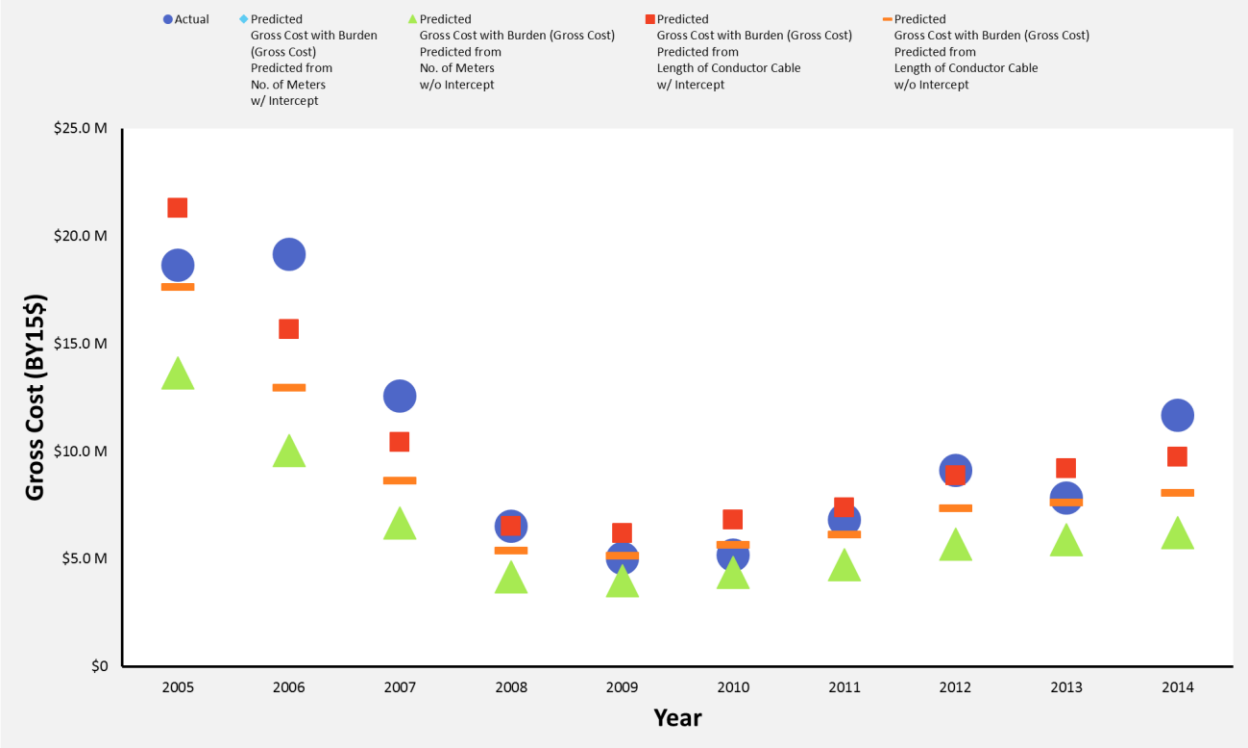

Figure 114. 2005 - 2015 Initial Validation Results for Gross Cost with Burden (Gross Cost) Models

16.2.6 Other Models: 2005 – 2015 Cross Validation Results

	Actual Cost	Predicted Cost: Predicted Gross Cost with Burden (Meters) Predicted from No. of Meters w/o Intercept	Predicted Cost: Predicted Gross Cost with Burden (Gross Cost) Predicted from Length of Conductor Cable w/o Intercept
2005 Removed	$18.6 M	$20.2 M	$24.8 M
2006 Removed	$19.2 M	$13.7 M	$17.0 M
2007 Removed	$12.6 M	$9.3 M	$11.1 M
2008 Removed	$6.5 M	$5.9 M	$7.4 M
2009 Removed	$5.0 M	$5.7 M	$7.0 M
2010 Removed	$5.2 M	$6.3 M	$7.7 M
2011 Removed	$6.8 M	$6.7 M	$8.3 M
2012 Removed	$9.1 M	$8.0 M	$9.7 M
2013 Removed	$7.8 M	$8.5 M	$10.2 M
2014 Removed	$11.7 M	$8.6 M	$10.3 M
2015 Removed	$4.6 M	$4.6 M	$5.5 M
Percent Error			
	2005 Removed	8%	33%
	2006 Removed	29%	11%
	2007 Removed	26%	12%
	2008 Removed	10%	13%
	2009 Removed	13%	38%
	2010 Removed	21%	48%
	2011 Removed	1%	22%
	2012 Removed	12%	7%
	2013 Removed	8%	31%
	2014 Removed	26%	12%
	2015 Removed	1%	20%
	Max % Error	29%	48%
	Avg % Error	14%	22%
	Mean Squared Error	5.13168E+12	6.05697E+12

Figure 115. Other Models: 2005 - 2015 Cross Validation Results

16.3 2008 – 2014 Data Set

16.3.1 Gross Cost with Burden (Equally Distributed)

16.3.1.1 Simple Linear Regression

16.3.1.1.1 Meters

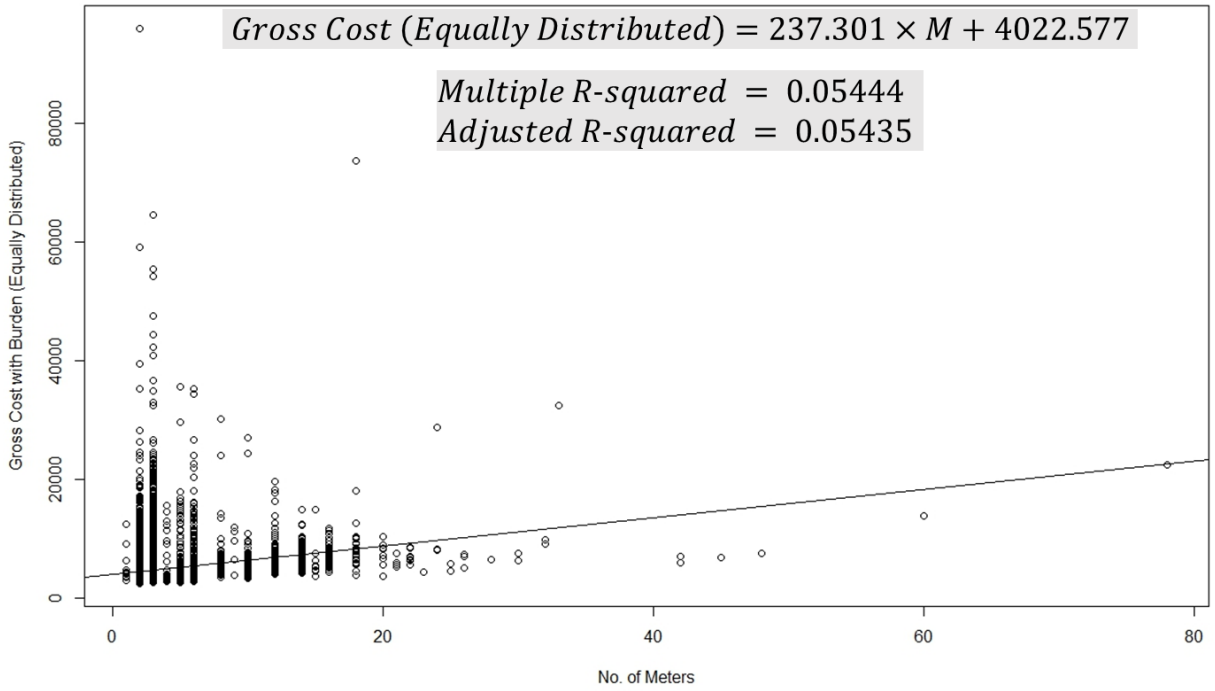

Figure 116. Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

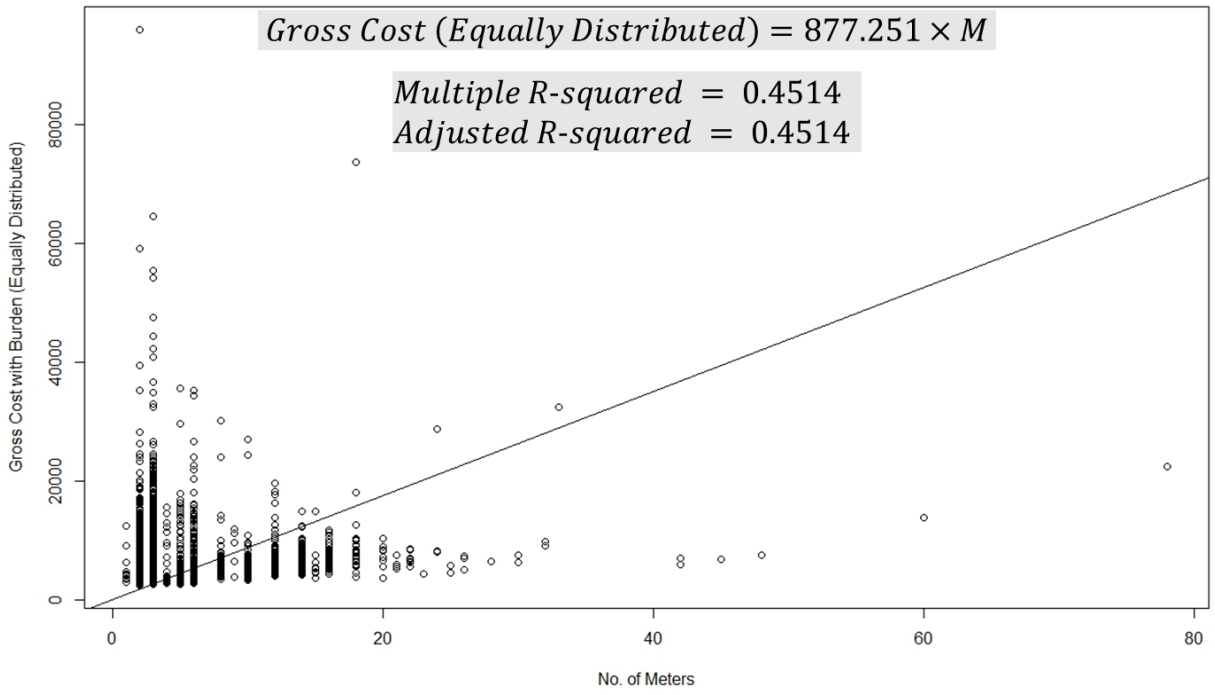

Figure 117. Linear Regression without Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.3.1.1.2 Length of Conductor Cable

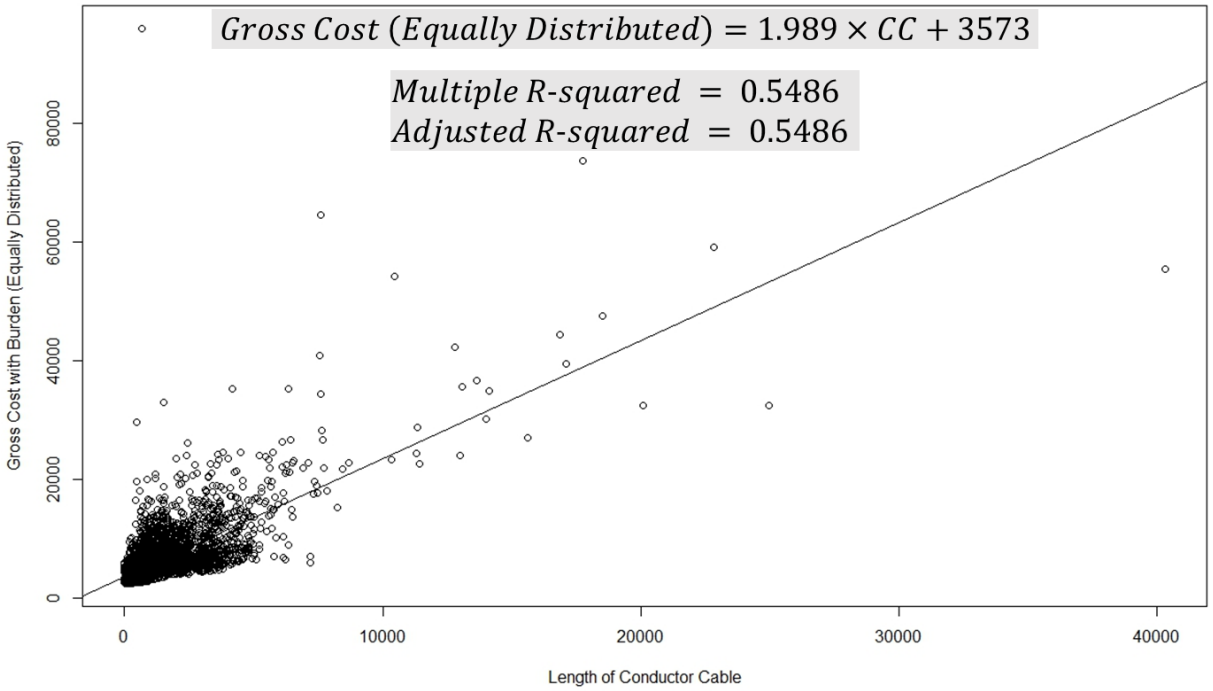

Figure 118. Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

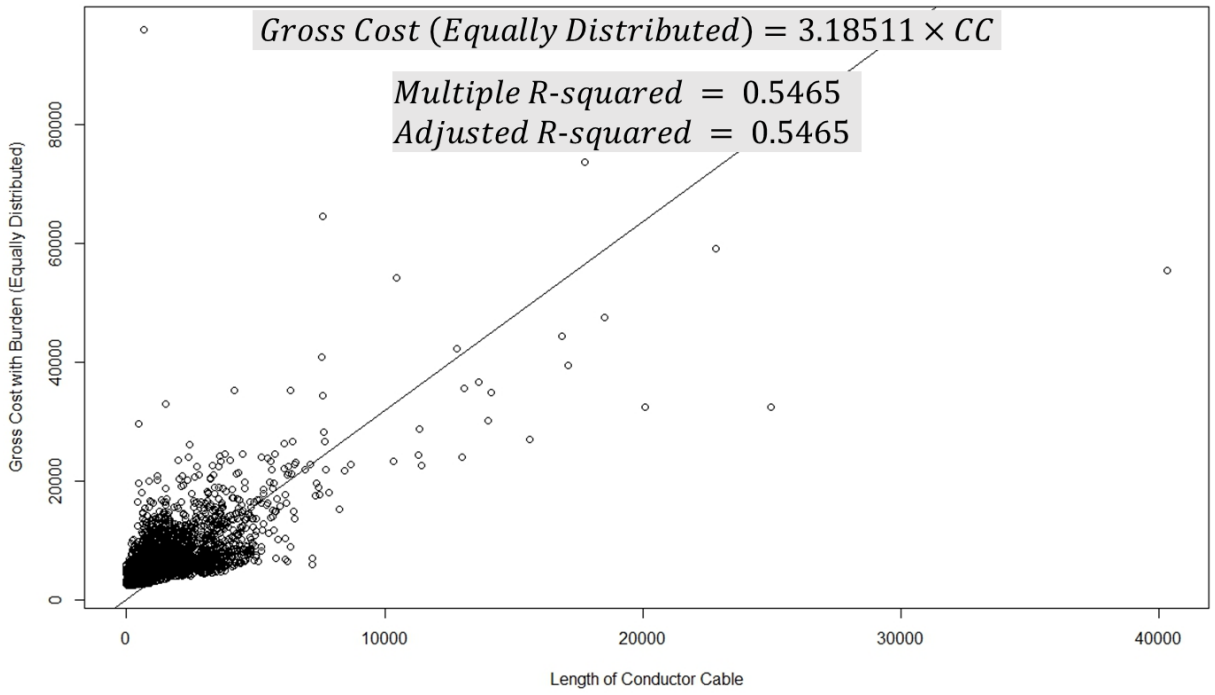

Figure 119. Linear Regression without Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.3.2 Gross Cost with Burden (Meters)

16.3.2.1 Simple Linear Regression

16.3.2.1.1 Meters

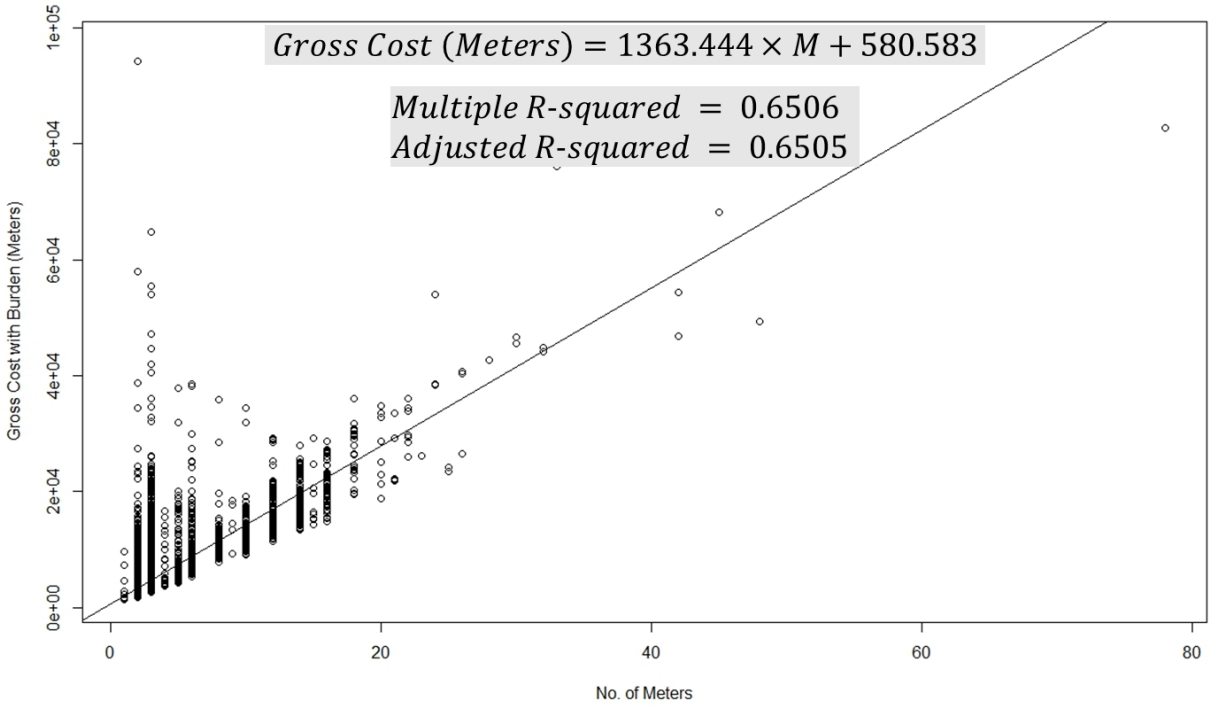

Figure 120. Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

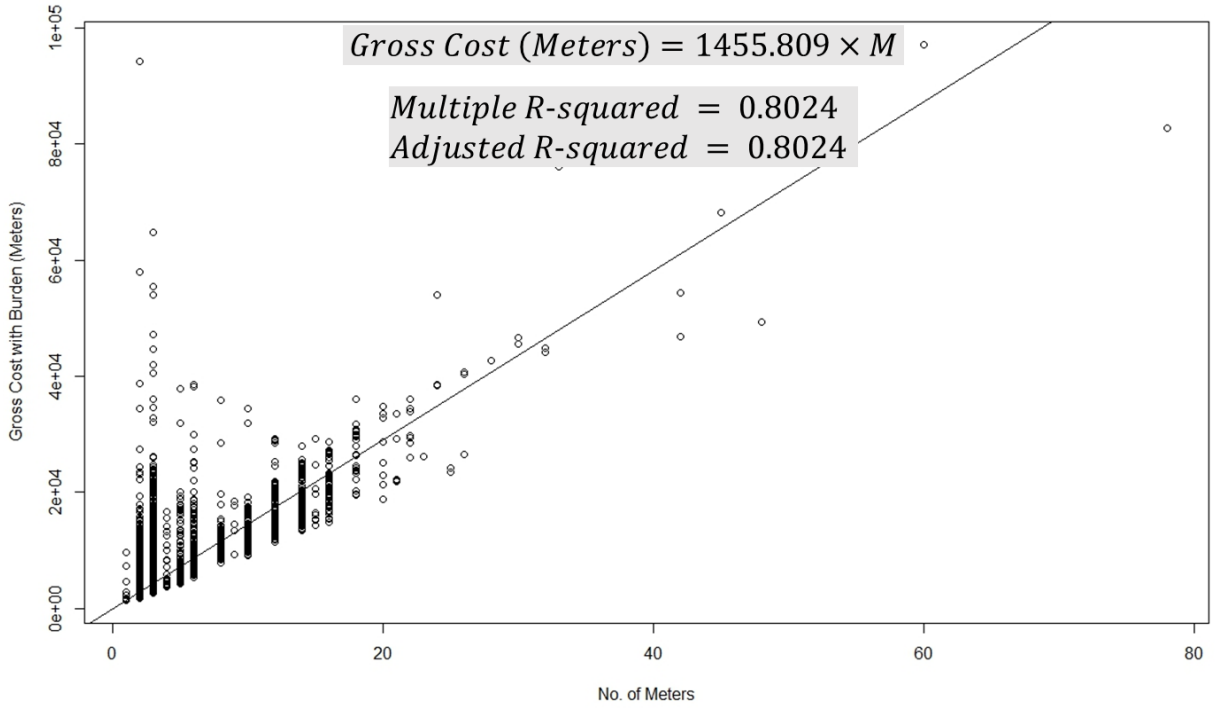

Figure 121. Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.2.1.2 Length of Conductor Cable

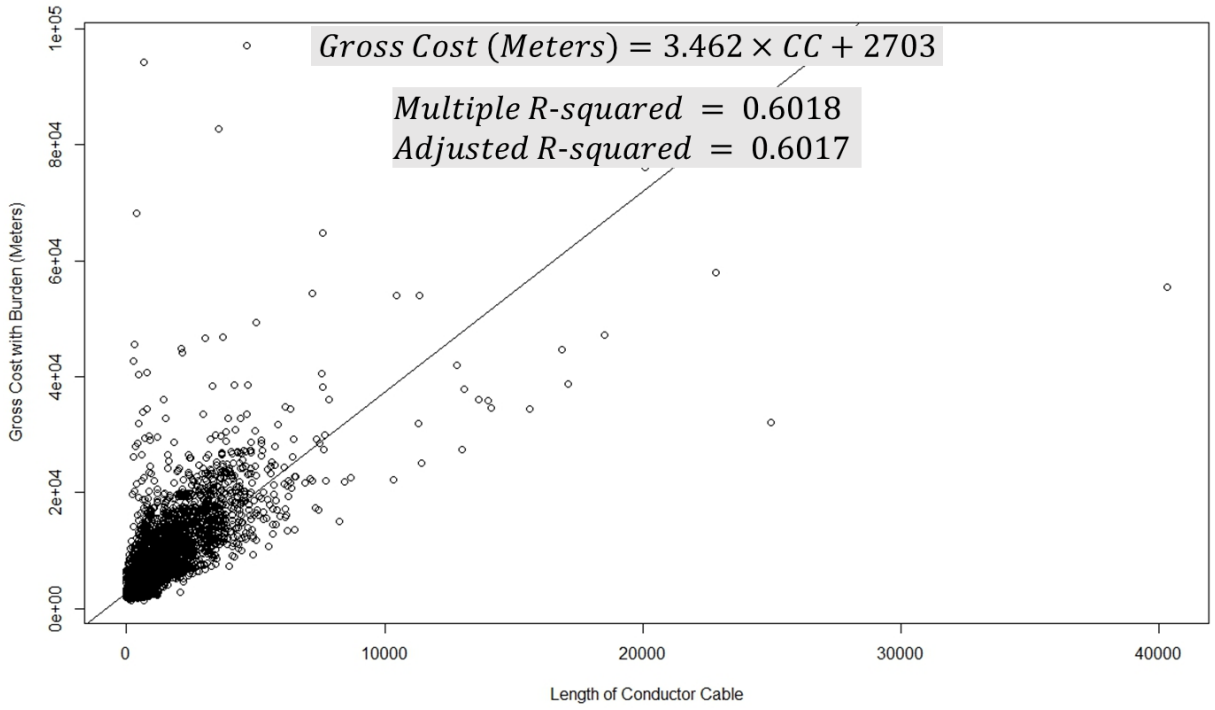

Figure 122. Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

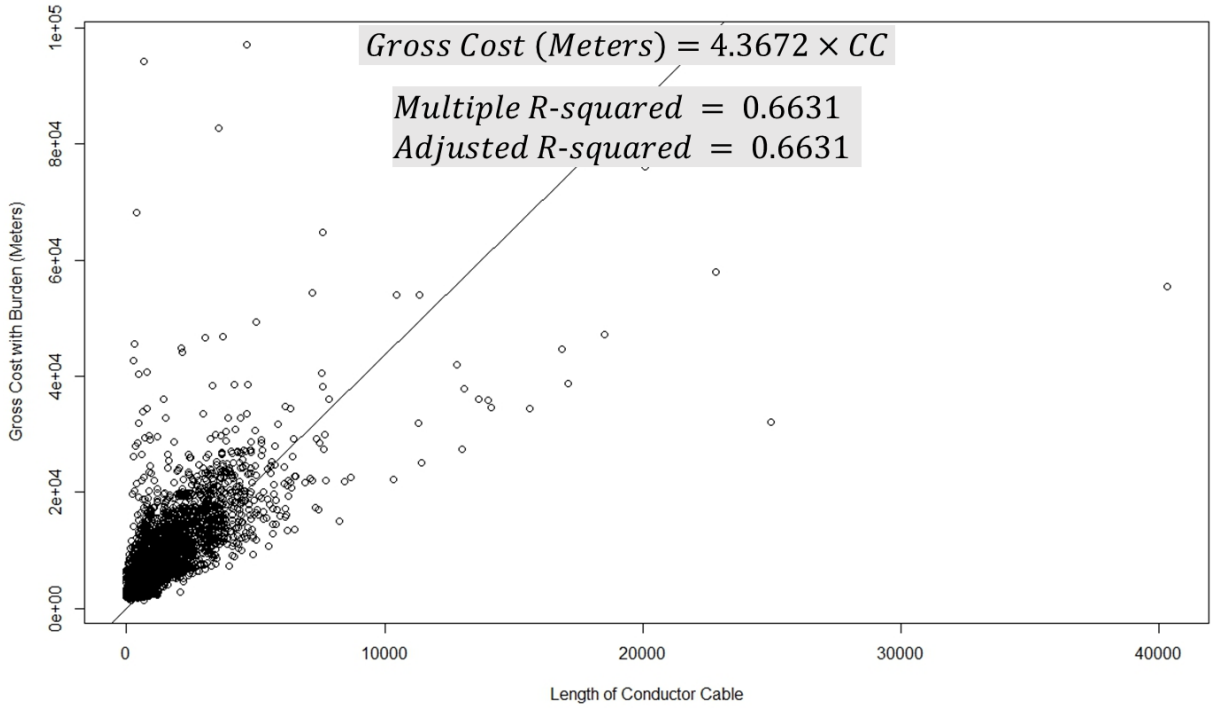

Figure 123. Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.3.3 Gross Cost with Burden (Gross Cost)

16.3.3.1 Simple Linear Regression

16.3.3.1.1 Meters

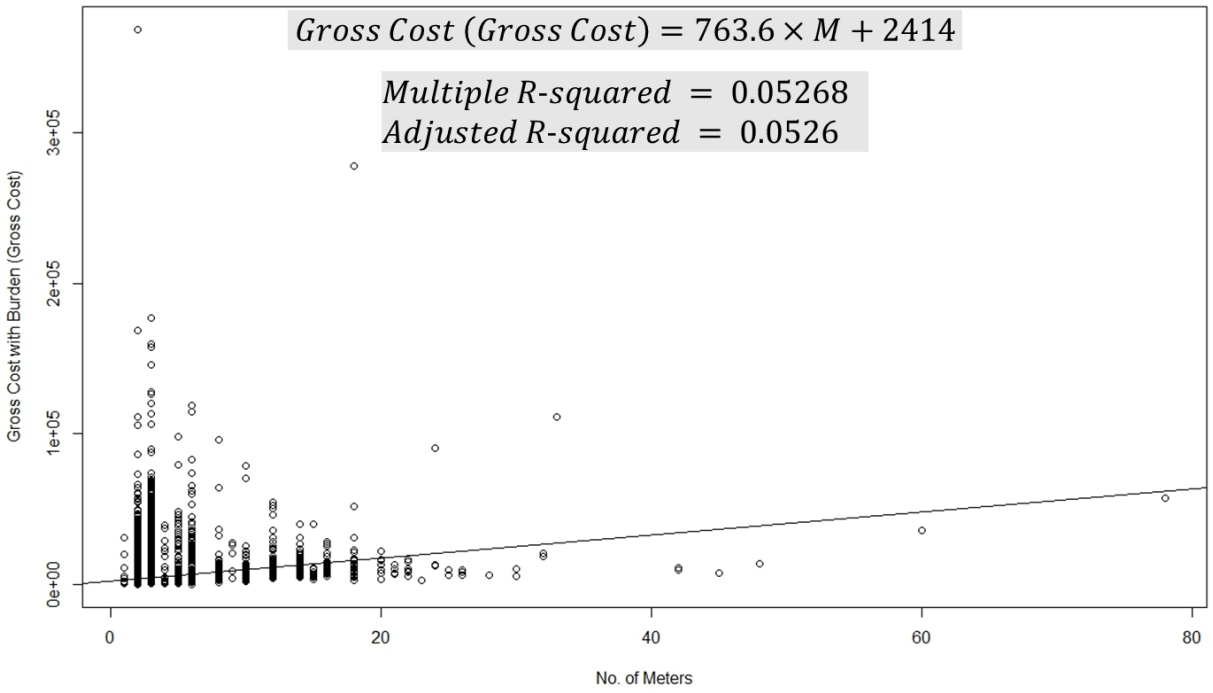

Figure 124. Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

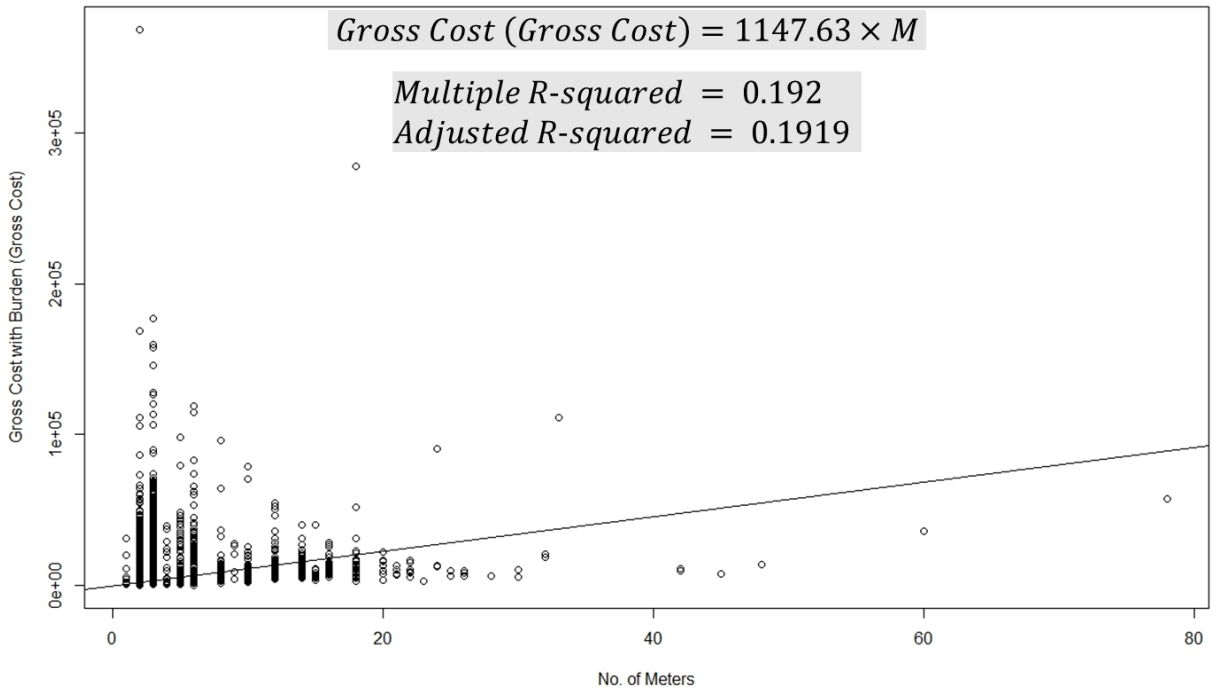

Figure 125. Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.3.3.1.2 Length of Conductor Cable

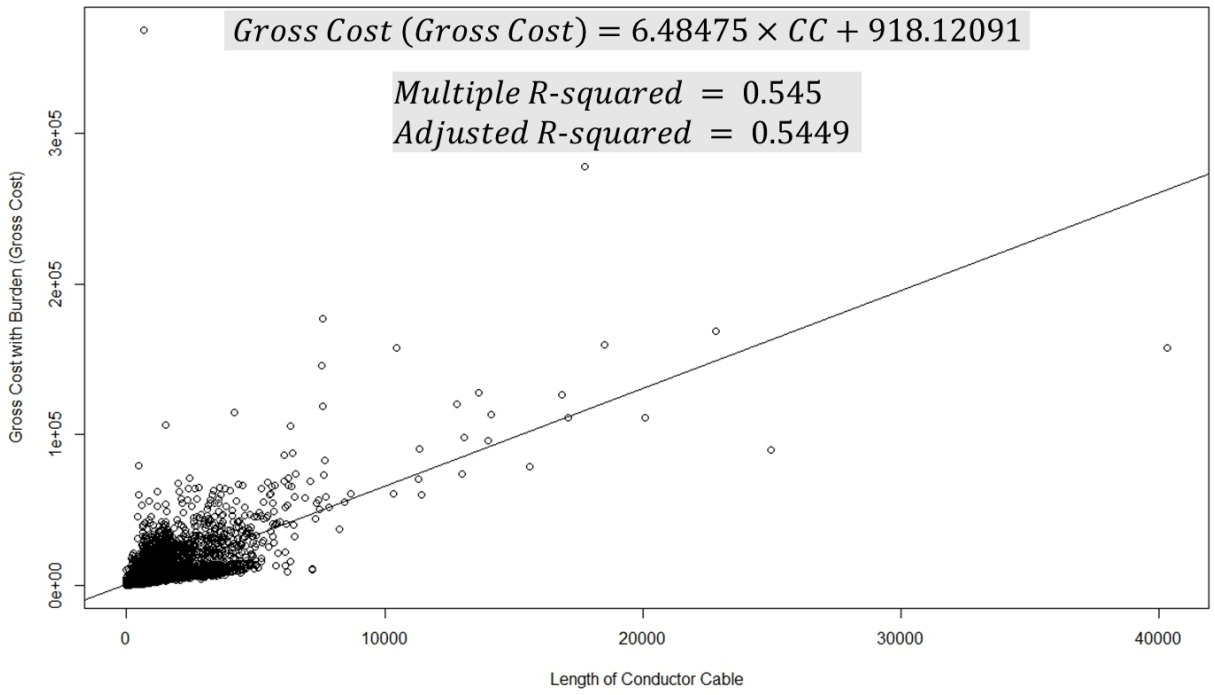

Figure 126. Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

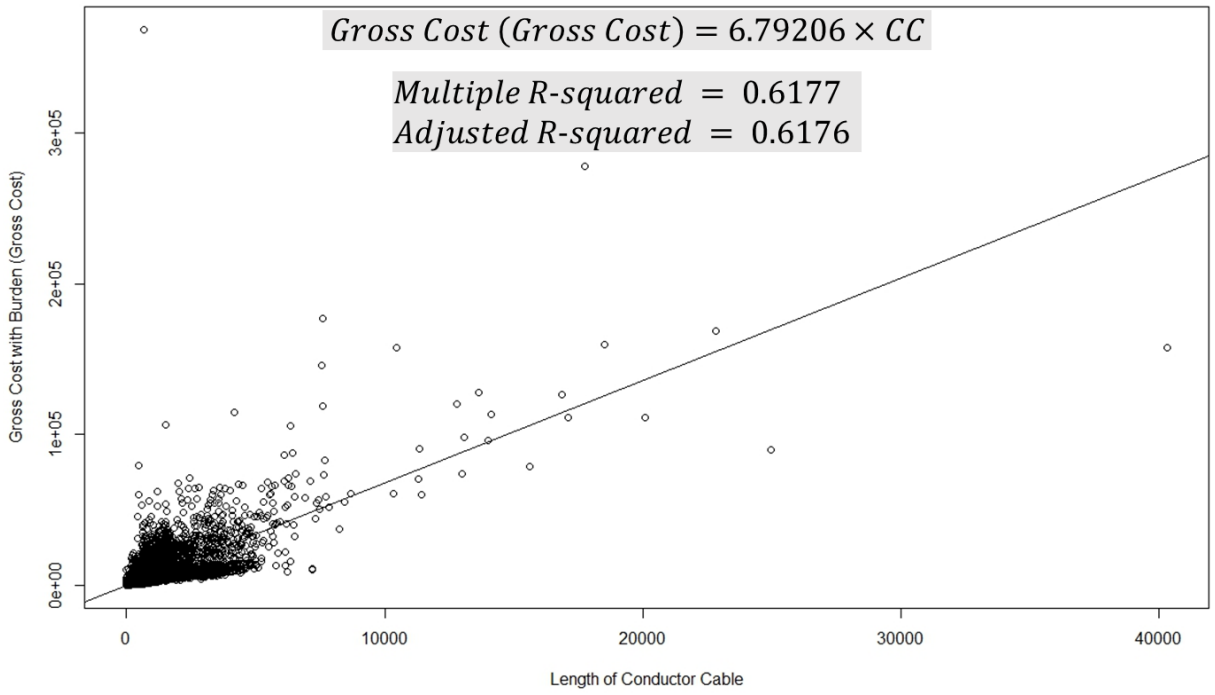

Figure 127. Linear Regression without Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.3.4 Cross Validation

16.3.4.1 Gross Cost with Burden (Equally Distributed)

16.3.4.1.1 Meters with Intercept

16.3.4.1.1.1 2008 Removed

```
Call:
lm(formula = TCE ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
 -7915  -1419   -924    619  91594

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4010.706     43.240   92.75   <2e-16 ***
M            234.157      9.713   24.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3080 on 9648 degrees of freedom
Multiple R-squared:  0.05682,   Adjusted R-squared:  0.05672
F-statistic: 581.2 on 1 and 9648 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 234.157 \times M + 4010.706$$

$$Multiple\ R-squared\ =\ 0.05682$$
$$Adjusted\ R-squared\ =\ 0.05672$$

Figure 128. 2008 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.3.4.1.1.2 2009 Removed

```
Call:
lm(formula = TCE ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
 -8071  -1247   -819    508  91443

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4161.347     44.388   93.75   <2e-16 ***
M            234.274      9.948   23.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3150 on 9646 degrees of freedom
Multiple R-squared:  0.05437,   Adjusted R-squared:  0.05427
F-statistic: 554.6 on 1 and 9646 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 234.274 \times M + 4161.347$$

$$Multiple\ R - squared\ = 0.05437$$
$$Adjusted\ R - squared\ = 0.05427$$

Figure 129. 2009 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.3.4.1.1.3 2010 Removed

```
Call:
lm(formula = TCE ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
 -7901  -1302   -869    483  91375

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4240.72      45.07   94.09   <2e-16 ***
M             228.36      10.19   22.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3138 on 9492 degrees of freedom
Multiple R-squared:  0.05028,   Adjusted R-squared:  0.05018
F-statistic: 502.5 on 1 and 9492 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 228.36 \times M + 4240.72$$

$$Multiple\ R-squared\ = 0.05028$$
$$Adjusted\ R-squared\ =\ 0.05018$$

Figure 130. 2010 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.3.4.1.1.4 2011 Removed

```
Call:
lm(formula = TCE ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
 -7880  -1521   -756    574  91492

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4112.73      45.98   89.45   <2e-16 ***
M             234.09      10.29   22.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3208 on 9371 degrees of freedom
Multiple R-squared:  0.05239,   Adjusted R-squared:  0.05228
F-statistic:   518 on 1 and 9371 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 234.09 \times M + 4112.73$$

$$Multiple\ R-squared\ = 0.05239$$
$$Adjusted\ R-squared\ =\ 0.05228$$

Figure 131. 2011 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.3.4.1.1.5 2012 Removed

```
Call:
lm(formula = TCE ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
 -7922  -1336   -863    741  91697

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3901.61      45.44   85.86   <2e-16 ***
M             236.91      10.54   22.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3134 on 9318 degrees of freedom
Multiple R-squared:  0.05147,   Adjusted R-squared:  0.05137
F-statistic: 505.7 on 1 and 9318 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 236.91 \times M + 3901.61$$

$$Multiple\ R - squared\ = 0.05147$$
$$Adjusted\ R - squared\ =\ 0.05137$$

Figure 132. 2012 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.3.4.1.1.6 2013 Removed

```
Call:
lm(formula = TCE ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
 -8736  -1494   -734    603  91531

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4035.45      45.62   88.46   <2e-16 ***
M             253.12      10.52   24.06   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3140 on 9231 degrees of freedom
Multiple R-squared:  0.05902,   Adjusted R-squared:  0.05892
F-statistic:   579 on 1 and 9231 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 253.12 \times M + 4035.45$$

$$Multiple\ R-squared\ = 0.05902$$
$$Adjusted\ R-squared\ =\ 0.05892$$

Figure 133. 2013 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.3.4.1.1.7 2014 Removed

```
Call:
lm(formula = TCE ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
 -7772  -1139   -667    188  60128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3690.16      42.51   86.81   <2e-16 ***
M             238.40      10.10   23.60   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2853 on 9184 degrees of freedom
Multiple R-squared:  0.05716,   Adjusted R-squared:  0.05706
F-statistic: 556.8 on 1 and 9184 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 238.40 \times M + 3690.16$$

$$Multiple\ R - squared\ = 0.05716$$
$$Adjusted\ R - squared\ = 0.05706$$

Figure 134. 2014 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from No. of Meters

16.3.4.1.2 Length of Conductor Cable with Intercept

16.3.4.1.2.1 2008 Removed

```
Call:
lm(formula = TCE ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-21856   -941   -343    734  91153

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.537e+03  2.517e+01   140.5   <2e-16 ***
CC          2.033e+00  1.982e-02   102.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2194 on 9648 degrees of freedom
Multiple R-squared:  0.5217,    Adjusted R-squared:  0.5216
F-statistic: 1.052e+04 on 1 and 9648 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 2.033 \times CC + 3537$$

$$Multiple\ R - squared\ = 0.5217$$
$$Adjusted\ R - squared\ = 0.5216$$

Figure 135. 2008 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.3.4.1.2.2 2009 Removed

```
Call:
lm(formula = TCE ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-27130   -829   -371    606  91049

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.692e+03  2.510e+01   147.1   <2e-16 ***
CC          1.957e+00  1.849e-02   105.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2203 on 9646 degrees of freedom
Multiple R-squared:  0.5373,    Adjusted R-squared:  0.5373
F-statistic: 1.12e+04 on 1 and 9646 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 1.957 \times CC + 3692$$

$$Multiple\ R - squared\ = 0.5373$$
$$Adjusted\ R - squared\ = 0.5373$$

Figure 136. 2009 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.3.4.1.2.3 2010 Removed

```
Call:
lm(formula = TCE ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-26321   -825   -412    557  91001

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.755e+03  2.468e+01   152.2   <2e-16 ***
CC          1.936e+00  1.786e-02   108.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2153 on 9492 degrees of freedom
Multiple R-squared:  0.5532,    Adjusted R-squared:  0.5532
F-statistic: 1.175e+04 on 1 and 9492 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 1.936 \times CC + 3755$$

$$Multiple\ R - squared\ = 0.5532$$
$$Adjusted\ R - squared\ = 0.5532$$

Figure 137. 2010 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.3.4.1.2.4 2011 Removed

```
Call:
lm(formula = TCE ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-28436  -1011   -295    680  91096

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.622e+03  2.556e+01   141.7   <2e-16 ***
CC          1.991e+00  1.867e-02   106.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2215 on 9371 degrees of freedom
Multiple R-squared:  0.5483,    Adjusted R-squared:  0.5482
F-statistic: 1.137e+04 on 1 and 9371 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 1.991 \times CC + 3622$$

$$Multiple\ R - squared\ = 0.5483$$
$$Adjusted\ R - squared\ = 0.5482$$

Figure 138. 2011 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.3.4.1.2.5 2012 Removed

```
Call:
lm(formula = TCE ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-28761   -856   -280    326  91260

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.450e+03  2.517e+01   137.0   <2e-16 ***
CC          2.004e+00  1.911e-02   104.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2179 on 9318 degrees of freedom
Multiple R-squared:  0.5412,    Adjusted R-squared:  0.5412
F-statistic: 1.099e+04 on 1 and 9318 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 2.004 \times CC + 3450$$

$$Multiple\ R - squared\ = 0.5412$$
$$Adjusted\ R - squared\ = 0.5412$$

Figure 139. 2012 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.3.4.1.2.6 2013 Removed

```
Call:
lm(formula = TCE ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-30319   -981   -355    662  91049

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.637e+03  2.516e+01   144.6   <2e-16 ***
CC          2.038e+00  1.924e-02   105.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2175 on 9231 degrees of freedom
Multiple R-squared:  0.5486,    Adjusted R-squared:  0.5486
F-statistic: 1.122e+04 on 1 and 9231 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 2.038 \times CC + 3637$$

$$Multiple\ R - squared\ = 0.5486$$
$$Adjusted\ R - squared\ = 0.5486$$

Figure 140. 2013 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.3.4.1.2.7 2014 Removed

```
Call:
lm(formula = TCE ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-27065   -710   -147    286  46279

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.311e+03  2.133e+01   155.2   <2e-16 ***
CC          1.965e+00  1.653e-02   118.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1844 on 9184 degrees of freedom
Multiple R-squared:  0.606,     Adjusted R-squared:  0.606
F-statistic: 1.413e+04 on 1 and 9184 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Equally\ Distributed) = 1.965 \times CC + 3311$$

$$Multiple\ R - squared\ = 0.606$$
$$Adjusted\ R - squared\ = 0.606$$

Figure 141. 2014 Removed, Linear Regression with Intercept: Gross Cost with Burden (Equally Distributed) predicted from Length of Conductor Cable

16.3.4.2 Gross Cost with Burden (Meters)

16.3.4.2.1 Meters with Intercept

16.3.4.2.1.1 2008 Removed

```
Call:
lm(formula = TCM ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-24064  -1052   -685    101  90851

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  550.724     43.751   12.59   <2e-16 ***
M           1362.959      9.828  138.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3117 on 9648 degrees of freedom
Multiple R-squared:  0.666,     Adjusted R-squared:  0.6659
F-statistic: 1.923e+04 on 1 and 9648 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1362.959 \times M + 550.724$$

$$Multiple\ R - squared\ = 0.666$$
$$Adjusted\ R - squared\ = 0.6659$$

Figure 142. 2008 Removed, Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.1.2 2009 Removed

```
Call:
lm(formula = TCM ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-25530  -1049   -658     11  90742

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   624.04      44.91   13.89   <2e-16 ***
M            1380.81      10.07  137.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3187 on 9646 degrees of freedom
Multiple R-squared:  0.6611,    Adjusted R-squared:  0.6611
F-statistic: 1.882e+04 on 1 and 9646 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1380.81 \times M + 624.04$$

$$Multiple\ R - squared\ = 0.6611$$
$$Adjusted\ R - squared\ = 0.6611$$

Figure 143. 2009 Removed, Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.1.3 2010 Removed

```
Call:
lm(formula = TCM ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-18505  -1044    -680    -15  90715

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   609.81      45.53   13.39   <2e-16 ***
M            1401.55      10.29  136.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3170 on 9492 degrees of freedom
Multiple R-squared:  0.6615,    Adjusted R-squared:  0.6615
F-statistic: 1.855e+04 on 1 and 9492 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1401.55 \times M + 609.81$$

$$Multiple\ R - squared\ = 0.6615$$
$$Adjusted\ R - squared\ = 0.6615$$

Figure 144. 2010 Removed, Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.1.4 2011 Removed

```
Call:
lm(formula = TCM ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-24709  -1110   -616     66  90795

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   591.22      46.56    12.7   <2e-16 ***
M            1370.70      10.42   131.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3250 on 9371 degrees of freedom
Multiple R-squared:  0.6488,    Adjusted R-squared:  0.6488
F-statistic: 1.731e+04 on 1 and 9371 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1370.70 \times M + 591.22$$

$$Multiple\ R - squared\ = 0.6488$$
$$Adjusted\ R - squared\ = 0.6488$$

Figure 145. 2011 Removed, Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.1.5 2012 Removed

```
Call:
lm(formula = TCM ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-22531  -1037   -700    130  90879

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   562.61      46.08   12.21   <2e-16 ***
M            1343.14      10.68  125.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3178 on 9318 degrees of freedom
Multiple R-squared:  0.6292,    Adjusted R-squared:  0.6291
F-statistic: 1.581e+04 on 1 and 9318 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1343.14 \times M + 562.61$$

$$Multiple\ R - squared\ = 0.6292$$
$$Adjusted\ R - squared\ = 0.6291$$

Figure 146. 2012 Removed, Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.1.6 2013 Removed

```
Call:
lm(formula = TCM ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-28544  -1085   -618     50  90784

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   501.89      45.73   10.98   <2e-16 ***
M            1421.02      10.54  134.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3148 on 9231 degrees of freedom
Multiple R-squared:  0.663,     Adjusted R-squared:  0.663
F-statistic: 1.816e+04 on 1 and 9231 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1421.02 \times M + 501.89$$

$$Multiple\ R - squared\ = 0.663$$
$$Adjusted\ R - squared\ =\ 0.663$$

Figure 147. 2013 Removed, Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.1.7 2014 Removed

```
Call:
lm(formula = TCM ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-14858   -933   -594   -231  60283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   670.80      42.61   15.74   <2e-16 ***
M            1243.39      10.13  122.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2860 on 9184 degrees of freedom
Multiple R-squared:  0.6214,    Adjusted R-squared:  0.6213
F-statistic: 1.507e+04 on 1 and 9184 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1243.39 \times M + 670.80$$

$$Multiple\ R - squared\ = 0.6214$$
$$Adjusted\ R - squared\ = 0.6213$$

Figure 148. 2014 Removed, Linear Regression with Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.2 Meters without Intercept

16.3.4.2.2.1 2008 Removed

```
Call:
lm(formula = TCM ~ M - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-30157   -680   -310    458  91231

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
M 1448.132      7.185   201.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3142 on 9649 degrees of freedom
Multiple R-squared:  0.8081,    Adjusted R-squared:  0.808
F-statistic: 4.062e+04 on 1 and 9649 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1448.132 \times M$$

$$Multiple\ R - squared\ = 0.8081$$
$$Adjusted\ R - squared\ =\ 0.808$$

Figure 149. 2008 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.2.2 2009 Removed

```
Call:
lm(formula = TCM ~ M - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-32449   -650   -255    424  91173

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
M 1477.517      7.344   201.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3219 on 9647 degrees of freedom
Multiple R-squared:  0.8075,    Adjusted R-squared:  0.8075
F-statistic: 4.048e+04 on 1 and 9647 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1477.517 \times M$$

$$Multiple\ R - squared\ = 0.8075$$
$$Adjusted\ R - squared\ = 0.8075$$

Figure 150. 2009 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.2.3 2010 Removed

```
Call:
lm(formula = TCM ~ M - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-22523   -657   -289    389  91132

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
M 1497.958      7.422   201.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3200 on 9493 degrees of freedom
Multiple R-squared:  0.811,     Adjusted R-squared:  0.811
F-statistic: 4.073e+04 on 1 and 9493 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1497.958 \times M$$

$$Multiple\ R - squared\ = 0.811$$
$$Adjusted\ R - squared\ = 0.811$$

Figure 151. 2010 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.2.4 2011 Removed

```
Call:
lm(formula = TCM ~ M - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-31268   -719   -238    456  91203

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
M 1462.369      7.573   193.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3277 on 9372 degrees of freedom
Multiple R-squared:  0.7992,    Adjusted R-squared:  0.7991
F-statistic: 3.729e+04 on 1 and 9372 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1462.369 \times M$$

$$Multiple\ R - squared\ = 0.7992$$
$$Adjusted\ R - squared\ = 0.7991$$

Figure 152. 2011 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.2.5 2012 Removed

```
Call:
lm(formula = TCM ~ M - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-29088   -671   -330    498  91259

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
M 1434.421      7.692   186.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3203 on 9319 degrees of freedom
Multiple R-squared:  0.7887,    Adjusted R-squared:  0.7887
F-statistic: 3.478e+04 on 1 and 9319 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1434.421 \times M$$

$$Multiple\ R - squared\ = 0.7887$$
$$Adjusted\ R - squared\ = 0.7887$$

Figure 153. 2012 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.2.6 2013 Removed

```
Call:
lm(formula = TCM ~ M - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-34340   -769   -297    380  91124

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
M 1501.762      7.603   197.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3168 on 9232 degrees of freedom
Multiple R-squared:  0.8087,    Adjusted R-squared:  0.8086
F-statistic: 3.902e+04 on 1 and 9232 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1501.762 \times M$$

$$Multiple\ R - squared\ = 0.8087$$
$$Adjusted\ R - squared\ =\ 0.8086$$

Figure 154. 2013 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.2.7 2014 Removed

```
Call:
lm(formula = TCM ~ M - 1, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-23066   -510   -181    194  60612

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
M 1357.217      7.185   188.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2898 on 9185 degrees of freedom
Multiple R-squared:  0.7953,    Adjusted R-squared:  0.7952
F-statistic: 3.568e+04 on 1 and 9185 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 1357.217 \times M$$

$$Multiple\ R - squared\ = 0.7953$$
$$Adjusted\ R - squared\ =\ 0.7952$$

Figure 155. 2014 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from No. of Meters

16.3.4.2.3 Length of Conductor Cable with Intercept

16.3.4.2.3.1 2008 Removed

```
Call:
lm(formula = TCM ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-64583   -954   -530    103  89044

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2514.0001    37.9716   66.21   <2e-16 ***
CC             3.7792     0.0299  126.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3309 on 9648 degrees of freedom
Multiple R-squared:  0.6234,    Adjusted R-squared:  0.6234
F-statistic: 1.597e+04 on 1 and 9648 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 3.7792 \times CC + 2514.0001$$

$$Multiple\ R - squared\ = 0.6234$$
$$Adjusted\ R - squared\ = 0.6234$$

Figure 156. 2008 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.3.4.2.3.2 2009 Removed

```
Call:
lm(formula = TCM ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-87609  -1084   -633     21  88996

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.765e+03  3.970e+01   69.63   <2e-16 ***
CC          3.481e+00  2.925e-02  118.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3485 on 9646 degrees of freedom
Multiple R-squared:  0.5948,    Adjusted R-squared:  0.5947
F-statistic: 1.416e+04 on 1 and 9646 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 3.481 \times CC + 2765$$

$$Multiple\ R - squared\ = 0.5948$$
$$Adjusted\ R - squared\ = 0.5947$$

Figure 157. 2009 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.3.4.2.3.3 2010 Removed

```
Call:
lm(formula = TCM ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-85203  -1066   -681     -8  88961

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.841e+03  3.937e+01   72.17   <2e-16 ***
CC          3.419e+00  2.848e-02  120.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3434 on 9492 degrees of freedom
Multiple R-squared:  0.6029,    Adjusted R-squared:  0.6028
F-statistic: 1.441e+04 on 1 and 9492 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 3.419 \times CC + 2841$$

$$Multiple\ R - squared\ = 0.6029$$
$$Adjusted\ R - squared\ = 0.6028$$

Figure 158. 2010 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.3.4.2.3.4 2011 Removed

```
Call:
lm(formula = TCM ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-85903  -1106   -709     69  89050

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.739e+03  4.048e+01   67.65   <2e-16 ***
CC          3.439e+00  2.957e-02  116.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3508 on 9371 degrees of freedom
Multiple R-squared:  0.5907,    Adjusted R-squared:  0.5907
F-statistic: 1.352e+04 on 1 and 9371 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 3.439 \times CC + 2739$$

$$Multiple\ R - squared\ = 0.5907$$
$$Adjusted\ R - squared\ = 0.5907$$

Figure 159. 2011 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.3.4.2.3.5 2012 Removed

```
Call:
lm(formula = TCM ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-83909   -999   -638    107  89180

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.641e+03  3.860e+01   68.43   <2e-16 ***
CC          3.392e+00  2.931e-02  115.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3342 on 9318 degrees of freedom
Multiple R-squared:  0.5898,    Adjusted R-squared:  0.5898
F-statistic: 1.34e+04 on 1 and 9318 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 3.392 \times CC + 2641$$

$$Multiple\ R - squared\ = 0.5898$$
$$Adjusted\ R - squared\ = 0.5898$$

Figure 160. 2012 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.3.4.2.3.6 2013 Removed

```
Call:
lm(formula = TCM ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-90620  -1082   -642      5  88939

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.771e+03  3.991e+01   69.42   <2e-16 ***
CC          3.555e+00  3.051e-02  116.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3450 on 9231 degrees of freedom
Multiple R-squared:  0.5953,    Adjusted R-squared:  0.5952
F-statistic: 1.358e+04 on 1 and 9231 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 3.555 \times CC + 2771$$

$$Multiple\ R - squared\ = 0.5953$$
$$Adjusted\ R - squared\ = 0.5952$$

Figure 161. 2013 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.3.4.2.3.7 2014 Removed

```
Call:
lm(formula = TCM ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-75197   -944   -592   -130  68791

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.635e+03  3.259e+01   80.87   <2e-16 ***
CC          3.176e+00  2.526e-02  125.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2817 on 9184 degrees of freedom
Multiple R-squared:  0.6326,    Adjusted R-squared:  0.6326
F-statistic: 1.581e+04 on 1 and 9184 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Meters) = 3.176 \times CC + 2635$$

$$Multiple\ R - squared\ = 0.6326$$
$$Adjusted\ R - squared\ = 0.6326$$

Figure 162. 2014 Removed, Linear Regression without Intercept: Gross Cost with Burden (Meters) predicted from Length of Conductor Cable

16.3.4.3 Gross Cost with Burden (Gross Cost)

16.3.4.3.1 Meters with Intercept

16.3.4.3.1.1 2008 Removed

```
Call:
lm(formula = TCG ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-28736  -2647  -2274  -1617 364621

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2406.78     143.24   16.80   <2e-16 ***
M             757.43      32.17   23.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10200 on 9648 degrees of freedom
Multiple R-squared:  0.05432,   Adjusted R-squared:  0.05422
F-statistic: 554.2 on 1 and 9648 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 757.43 \times M + 2406.78$$

$$Multiple\ R - squared\ = 0.05432$$
$$Adjusted\ R - squared\ = 0.05422$$

Figure 163. 2008 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.3.4.3.1.2 2009 Removed

```
Call:
lm(formula = TCG ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-29180  -2747  -2361  -1697 364488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2524.97     147.29   17.14   <2e-16 ***
M             764.67      33.01   23.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10450 on 9646 degrees of freedom
Multiple R-squared:  0.0527,    Adjusted R-squared:  0.0526
F-statistic: 536.6 on 1 and 9646 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 764.67 \times M + 2524.97$$

$$Multiple\ R - squared\ = 0.0527$$
$$Adjusted\ R - squared\ = 0.0526$$

Figure 164. 2009 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.3.4.3.1.3 2010 Removed

```
Call:
lm(formula = TCG ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-29092  -2778  -2405  -1750 364429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2591.5      151.3   17.13   <2e-16 ***
M              761.2       34.2   22.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10540 on 9492 degrees of freedom
Multiple R-squared:  0.0496,    Adjusted R-squared:  0.0495
F-statistic: 495.4 on 1 and 9492 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 761.2 \times M + 2591.5$$

$$Multiple\ R - squared\ = 0.0496$$
$$Adjusted\ R - squared\ = 0.0495$$

Figure 165. 2010 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.3.4.3.1.4 2011 Removed

```
Call:
lm(formula = TCG ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-28793  -2732  -2344  -1660 364536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2493.34     150.76   16.54   <2e-16 ***
M             756.77      33.73   22.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10520 on 9371 degrees of freedom
Multiple R-squared:  0.05099,   Adjusted R-squared:  0.05089
F-statistic: 503.5 on 1 and 9371 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 756.77 \times M + 2493.34$$

$$Multiple\ R - squared\ = 0.05099$$
$$Adjusted\ R - squared\ = 0.05089$$

Figure 166. 2011 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.3.4.3.1.5 2012 Removed

```
Call:
lm(formula = TCG ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-28231  -2596  -2260  -1601 364687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2361.33     146.61   16.11   <2e-16 ***
M             747.22      33.99   21.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10110 on 9318 degrees of freedom
Multiple R-squared:  0.0493,    Adjusted R-squared:  0.0492
F-statistic: 483.2 on 1 and 9318 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 747.22 \times M + 2361.33$$

$$Multiple\ R - squared\ = 0.0493$$
$$Adjusted\ R - squared\ = 0.0492$$

Figure 167. 2012 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.3.4.3.1.6 2013 Removed

```
Call:
lm(formula = TCG ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-31357  -2654  -2291  -1653 364580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2328.1      150.9   15.43   <2e-16 ***
M              817.4       34.8   23.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10390 on 9231 degrees of freedom
Multiple R-squared:  0.0564,    Adjusted R-squared:  0.0563
F-statistic: 551.8 on 1 and 9231 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 817.4 \times M + 2328.1$$

$$Multiple\ R - squared\ = 0.0564$$
$$Adjusted\ R - squared\ = 0.0563$$

Figure 168. 2013 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.3.4.3.1.7 2014 Removed

```
Call:
lm(formula = TCG ~ M, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-23818  -2410  -2093  -1582 172682

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2195.92     131.94   16.64   <2e-16 ***
M             735.75      31.36   23.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8854 on 9184 degrees of freedom
Multiple R-squared:  0.05656,   Adjusted R-squared:  0.05645
F-statistic: 550.6 on 1 and 9184 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 735.75 \times M + 2195.92$$

$$Multiple\ R - squared\ = 0.05656$$
$$Adjusted\ R - squared\ = 0.05645$$

Figure 169. 2014 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from No. of Meters

16.3.4.3.2 Length of Conductor Cable with Intercept

16.3.4.3.2.1 2008 Removed

```
Call:
lm(formula = TCG ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-78758  -1020   -167    284 363180

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 775.23964   83.00475    9.34   <2e-16 ***
CC            6.74660    0.06536  103.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7234 on 9648 degrees of freedom
Multiple R-squared:  0.5248,    Adjusted R-squared:  0.5247
F-statistic: 1.065e+04 on 1 and 9648 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 6.74660 \times CC + 775.23964$$

$$Multiple\ R - squared\ = 0.5248$$
$$Adjusted\ R - squared\ = 0.5247$$

Figure 170. 2008 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.3.4.3.2.2 2009 Removed

```
Call:
lm(formula = TCG ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
    Min      1Q  Median      3Q     Max
-103771   -1125    -307     204  363199

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 948.0111    83.5926   11.34   <2e-16 ***
CC            6.4641     0.0616  104.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7338 on 9646 degrees of freedom
Multiple R-squared:  0.5331,    Adjusted R-squared:  0.533
F-statistic: 1.101e+04 on 1 and 9646 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 6.4641 \times CC + 948.0111$$

$$Multiple\ R - squared\ = 0.5331$$
$$Adjusted\ R - squared\ = 0.533$$

Figure 171. 2009 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.3.4.3.2.3 2010 Removed

```
Call:
lm(formula = TCG ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
    Min      1Q  Median      3Q     Max
-102669   -1163    -325     186  363183

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 983.02191   83.77519   11.73   <2e-16 ***
CC            6.43587    0.06062  106.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7307 on 9492 degrees of freedom
Multiple R-squared:  0.5429,    Adjusted R-squared:  0.5428
F-statistic: 1.127e+04 on 1 and 9492 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 6.43587 \times CC + 983.02191$$

$$Multiple\ R - squared\ = 0.5429$$
$$Adjusted\ R - squared\ = 0.5428$$

Figure 172. 2010 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.3.4.3.2.4 2011 Removed

```
Call:
lm(formula = TCG ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
    Min      1Q  Median      3Q     Max
-105333   -1108    -274     268  363252

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 867.70445   84.08710   10.32   <2e-16 ***
CC            6.50483    0.06143  105.90   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7287 on 9371 degrees of freedom
Multiple R-squared:  0.5448,    Adjusted R-squared:  0.5447
F-statistic: 1.121e+04 on 1 and 9371 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 6.50483 \times CC + 867.70445$$

$$Multiple\ R - squared\ = 0.5448$$
$$Adjusted\ R - squared\ = 0.5447$$

Figure 173. 2011 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.3.4.3.2.5 2012 Removed

```
Call:
lm(formula = TCG ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
    Min      1Q  Median      3Q     Max
-100822   -1041    -291      98  363302

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 893.75925   82.07376   10.89   <2e-16 ***
CC            6.39227    0.06232  102.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7107 on 9318 degrees of freedom
Multiple R-squared:  0.5303,    Adjusted R-squared:  0.5303
F-statistic: 1.052e+04 on 1 and 9318 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 6.39227 \times CC + 893.75925$$

$$Multiple\ R - squared\ = 0.5303$$
$$Adjusted\ R - squared\ = 0.5303$$

Figure 174. 2012 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.3.4.3.2.6 2013 Removed

```
Call:
lm(formula = TCG ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
    Min      1Q  Median      3Q     Max
-113512   -1033    -352     147  363011

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 971.75619   83.52412   11.63   <2e-16 ***
CC            6.70515    0.06386  104.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7220 on 9231 degrees of freedom
Multiple R-squared:  0.5442,    Adjusted R-squared:  0.5442
F-statistic: 1.102e+04 on 1 and 9231 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 6.70515 \times CC + 971.75619$$

$$Multiple\ R - squared\ = 0.5442$$
$$Adjusted\ R - squared\ = 0.5442$$

Figure 175. 2013 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.3.4.3.2.7 2014 Removed

```
Call:
lm(formula = TCG ~ CC, data = ReducedDataSetMiscodedMetersRemoved)

Residuals:
   Min     1Q Median     3Q    Max
-91291   -919   -325     19 129317

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 975.31411   65.24205   14.95   <2e-16 ***
CC            6.15379    0.05056  121.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5639 on 9184 degrees of freedom
Multiple R-squared:  0.6173,    Adjusted R-squared:  0.6172
F-statistic: 1.481e+04 on 1 and 9184 DF,  p-value: < 2.2e-16
```

$$Gross\ Cost\ (Gross\ Cost) = 6.15379 \times CC + 975.31411$$

$$Multiple\ R - squared\ = 0.6173$$
$$Adjusted\ R - squared\ = 0.6172$$

Figure 176. 2014 Removed, Linear Regression with Intercept: Gross Cost with Burden (Gross Cost) predicted from Length of Conductor Cable

16.3.5 Other Models: 2008 – 2014 Initial Validation Results

		No. of Meters					
	Actual Gross Cost	Predicted Gross Cost with Burden (Equally Distributed) Predicted from No. of Meters w/ Intercept	Predicted Gross Cost with Burden (Equally Distributed) Predicted from No. of Meters w/o Intercept	Predicted Gross Cost with Burden (Meters) Predicted from No. of Meters w/ Intercept	Predicted Gross Cost with Burden (Meters) Predicted from No. of Meters w/o Intercept	Predicted Gross Cost with Burden (Gross Cost) Predicted from No. of Meters w/ Intercept	Predicted Gross Cost with Burden (Gross Cost) Predicted from No. of Meters w/o Intercept
2008	$6.5 M	$6.2 M	$3.5 M	$6.2 M	$5.8 M	$6.2 M	$4.6 M
2009	$5.0 M	$5.9 M	$3.3 M	$5.9 M	$5.5 M	$5.9 M	$4.4 M
2010	$5.2 M	$6.5 M	$3.7 M	$6.5 M	$6.1 M	$6.5 M	$4.8 M
2011	$6.8 M	$7.0 M	$4.0 M	$7.0 M	$6.6 M	$7.0 M	$5.2 M
2012	$9.1 M	$8.5 M	$4.8 M	$8.5 M	$7.9 M	$8.5 M	$6.2 M
2013	$7.8 M	$8.8 M	$4.9 M	$8.8 M	$8.2 M	$8.8 M	$6.5 M
2014	$11.7 M	$9.3 M	$5.2 M	$9.3 M	$8.7 M	$9.3 M	$6.9 M
		Percent Error					
	2008	5%	46%	5%	11%	5%	30%
	2009	18%	34%	18%	10%	18%	13%
	2010	26%	29%	26%	18%	26%	7%
	2011	3%	42%	3%	3%	3%	24%
	2012	7%	48%	7%	13%	7%	32%
	2013	12%	37%	12%	5%	12%	17%
	2014	21%	55%	21%	25%	21%	41%
	Max % Error	26%	55%	26%	25%	26%	41%
	Avg % Error	13%	41%	13%	12%	13%	23%
	Mean Squared Error	1.39E+12	1.30E+13	1.39E+12	1.73E+12	1.39E+12	5.75E+12

Figure 177. 2008 - 2014 Initial Model Validation Results predicted from No. of Meters

		Length of Coductor Cable					
	Actual Gross Cost	Predicted Gross Cost with Burden (Equally Distributed) Predicted from Length of Conductor Cable w/ Intercept	Predicted Gross Cost with Burden (Equally Distributed) Predicted from Length of Conductor Cable w/o Intercept	Predicted Gross Cost with Burden (Meters) Predicted from Length of Conductor Cable w/ Intercept	Predicted Gross Cost with Burden (Meters) Predicted from Length of Conductor Cable w/o Intercept	Predicted Gross Cost with Burden (Gross Cost) Predicted from Length of Conductor Cable w/ Intercept	Predicted Gross Cost with Burden (Gross Cost) Predicted from Length of Conductor Cable w/o Intercept
2008	$6.5 M	$6.2 M	$2.5 M	$6.2 M	$3.4 M	$6.2 M	$5.2 M
2009	$5.0 M	$5.9 M	$2.3 M	$5.9 M	$3.2 M	$5.9 M	$5.0 M
2010	$5.2 M	$6.5 M	$2.6 M	$6.5 M	$3.5 M	$6.5 M	$5.5 M
2011	$6.8 M	$7.0 M	$2.8 M	$7.0 M	$3.8 M	$7.0 M	$5.9 M
2012	$9.1 M	$8.5 M	$3.3 M	$8.5 M	$4.6 M	$8.5 M	$7.1 M
2013	$7.8 M	$8.8 M	$3.5 M	$8.8 M	$4.8 M	$8.8 M	$7.4 M
2014	$11.7 M	$9.3 M	$3.7 M	$9.3 M	$5.0 M	$9.3 M	$7.8 M
		Percent Error					
	2008	5%	62%	5%	48%	5%	20%
	2009	18%	53%	18%	36%	18%	1%
	2010	26%	50%	26%	32%	26%	6%
	2011	3%	59%	3%	44%	3%	13%
	2012	7%	63%	7%	50%	7%	22%
	2013	12%	56%	12%	39%	12%	5%
	2014	21%	69%	21%	57%	21%	33%
	Max % Error	26%	69%	26%	57%	26%	33%
	Avg % Error	13%	59%	13%	44%	13%	14%
	Mean Squared Error	1.39E+12	2.33E+13	1.39E+12	1.41E+13	1.39E+12	3.04E+12

Figure 178. 2008 - 2014 Initial Model Validation Results predicted from Length of Conductor Cable

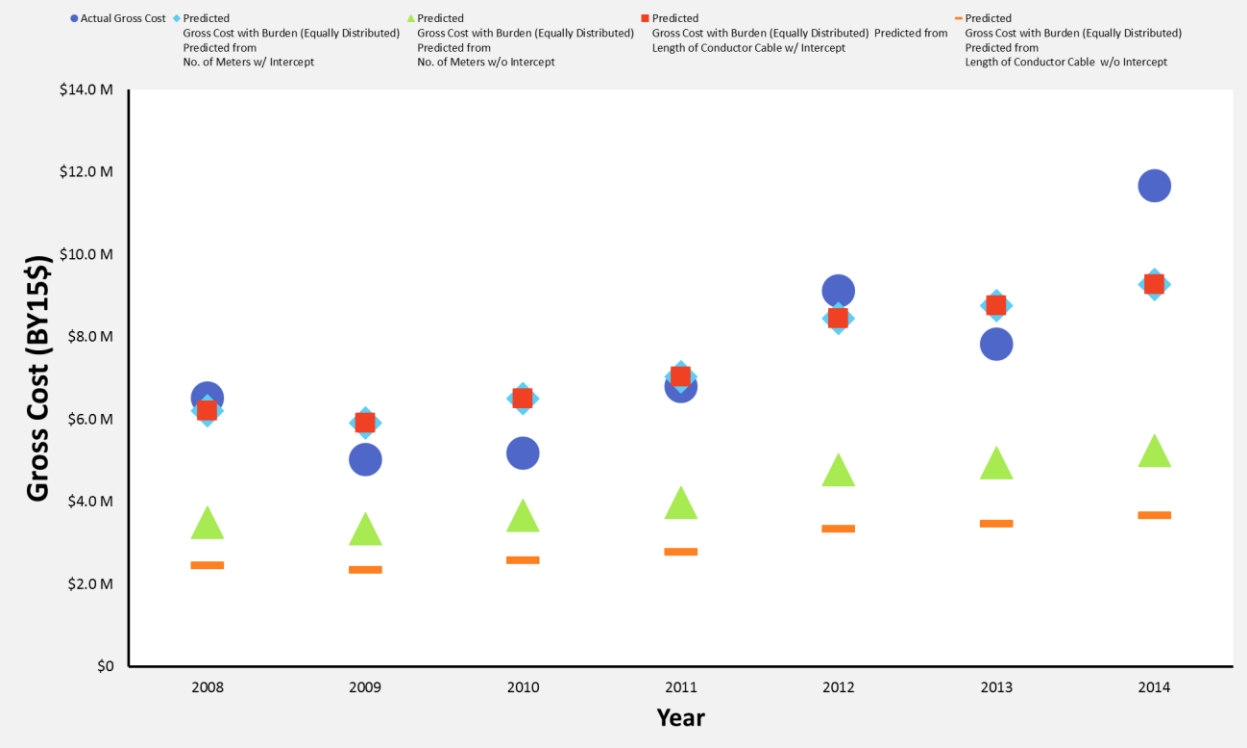

Figure 179. 2008 - 2014 Initial Validation Results for Gross Cost with Burden (Equally Distributed) Models

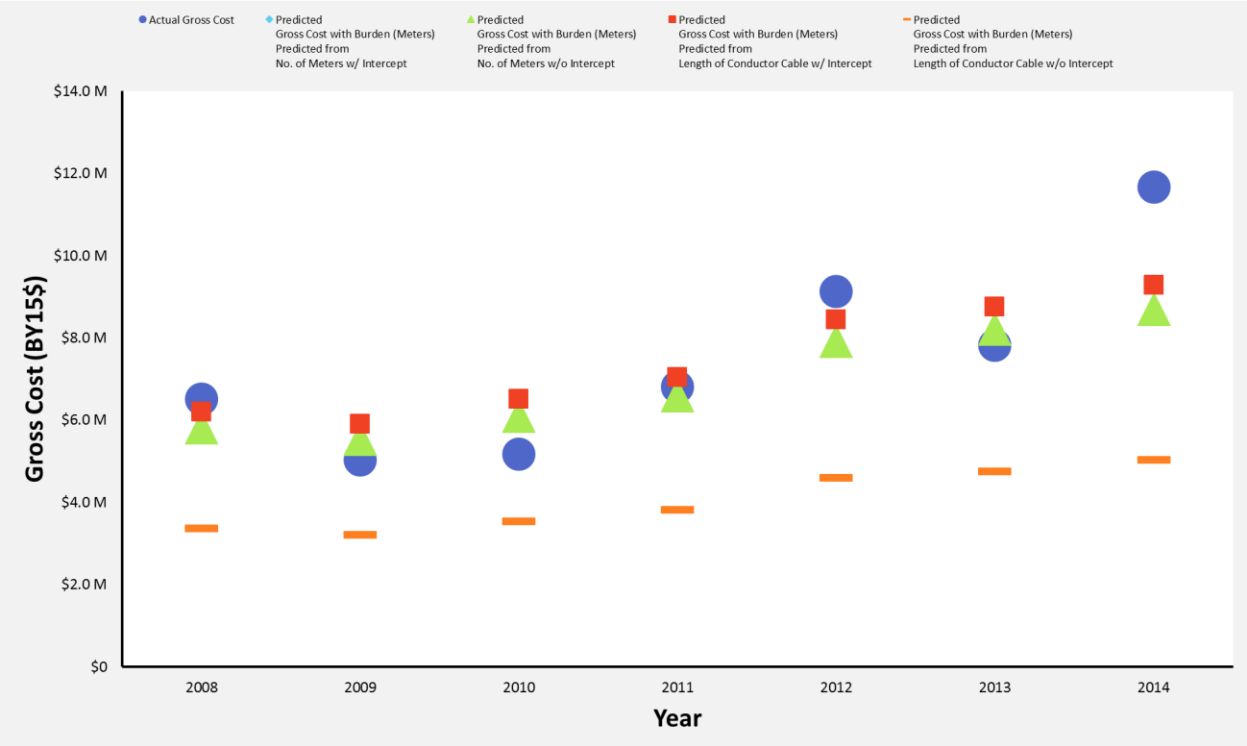

Figure 180. 2008 - 2014 Initial Validation Results for Gross Cost with Burden (Meters) Models

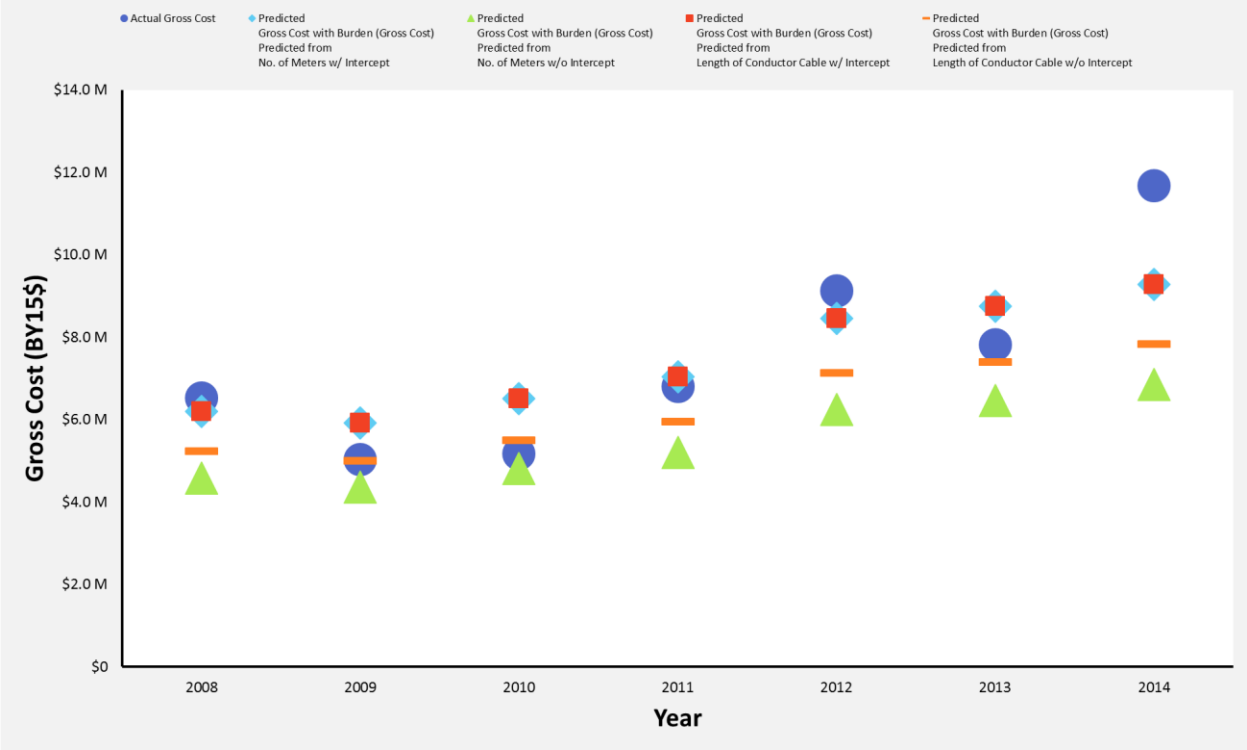

Figure 181. 2008 - 2014 Initial Validation Results for Gross Cost with Burden (Gross Cost) Models

16.3.6 Other Models: 2008 – 2014 Cross Validation Results

		METERS				CONDUCTOR CABLE		
	Actual Gross Cost	Predicted Gross Cost with Burden (Equally Distributed) w/ Intercept	Predicted Gross Cost with Burden (Meters) w/ Intercept	Predicted Gross Cost with Burden (Meters) w/o Intercept	Predicted Gross Cost with Burden (Gross Cost) w/ Intercept	Predicted Gross Cost with Burden (Equally Distributed) w/ Intercept	Predicted Gross Cost with Burden (Meters) w/ Intercept	Predicted Gross Cost with Burden (Gross Cost) w/ Intercept
2008 Removed	$6.5 M	$6.2 M	$6.2 M	$5.8 M	$6.2 M	$6.1 M	$6.2 M	$6.2 M
2009 Removed	$5.0 M	$6.0 M	$6.0 M	$5.6 M	$6.0 M	$6.0 M	$6.0 M	$6.0 M
2010 Removed	$5.2 M	$6.7 M	$6.7 M	$6.3 M	$6.7 M	$6.7 M	$6.7 M	$6.7 M
2011 Removed	$6.8 M	$7.1 M	$7.1 M	$6.6 M	$7.1 M	$7.1 M	$7.1 M	$7.1 M
2012 Removed	$9.1 M	$8.3 M	$8.3 M	$7.8 M	$8.3 M	$8.3 M	$8.3 M	$8.3 M
2013 Removed	$7.8 M	$8.9 M	$8.9 M	$8.5 M	$8.9 M	$8.9 M	$8.9 M	$8.9 M
2014 Removed	$11.7 M	$8.8 M	$8.8 M	$8.1 M	$8.8 M	$8.8 M	$8.8 M	$8.8 M
Percent Error								
	2008 Removed	6%	6%	11%	6%	6%	6%	6%
	2009 Removed	20%	20%	12%	20%	20%	20%	20%
	2010 Removed	29%	29%	21%	29%	29%	29%	29%
	2011 Removed	4%	4%	3%	4%	4%	4%	4%
	2012 Removed	9%	9%	14%	9%	9%	9%	9%
	2013 Removed	14%	14%	8%	14%	14%	14%	14%
	2014 Removed	25%	25%	31%	25%	25%	25%	25%
	Max % Error	29%	29%	31%	29%	29%	29%	29%
	Avg % Error	17%	17%	16%	17%	17%	17%	17%
	Mean Squared Error	3.77E+12	3.77E+12	1.07E+13	3.77E+12	3.79E+12	3.77E+12	3.77E+12

Figure 182. Other Models: 2008 - 2014 Cross Validation Results