Isolating and quantifying factors affecting body and paint business for Volvo Cars

Muhammad Awais Khan
Abstract
This thesis focuses on identifying the degree of contribution of the most important factors affecting Body and Paint business of Volvo Car Corporation in Sweden. It is clear that Body and Paint business for VCCS directly depends on the number of registered accidents. Our major purpose is to determine the factors which have direct or indirect effect on reduction in the number of accidents in Sweden and to analyze in which degree they may affect the business. During the interviews with senior staff members, we discover that particularly city safety cars are mentioned by most of the specialists. Other important factors highlighted were mileage, weather, company car/ private car and age of a car.

City Safety is a technology designed to help the driver mitigate, and in certain situations avoid, collisions at low speed by automatically braking the vehicle. The estimated claim rate frequency i.e. claims per contract rate was 50% lower for city safety equipped; then other warranty cars models without system. The study also analysis the effect of rain, mean temperature and snow on Volvo Body part sales in Stockholm Sweden. Temperature snow impacted road accidents significantly. Snow was shown to be the leading variable, as the number of accidents increases sharply with increased snowfall. Temperature is the second important variable in the list i.e. as the temperature decreases by 1°C the sales of body and paint business in Stockholm increases by 1.6%.

Time variable such as weekday, month, and year also plays significant role in this model. During Fridays 51% high accidents are expected then accidents occurred on Sundays.
Acknowledgements

I would like to thank everyone who helped me during the development of this thesis paper; in particular, my supervisor Prof. Anders Nordgaard. I would also like to thank Mattias Villani and Oleg for providing guidance and support, as well as giving me suggestions for improvements.

I would also like to thank Bart Smits for introducing such an interesting topic and providing me an opportunity to work with it. His comments and suggestions were very helpful in researching and understanding the business. I would also like to thank Mikael Thorin; Volvia, Gothenburg for providing us the necessary research data.
# Table of contents

1. Introduction ............................................................................................................................................. 6  
2. Background ............................................................................................................................................... 6  
3. Objective .................................................................................................................................................. 8  
4. Outline of Thesis ....................................................................................................................................... 8  
5. Problem discussion .................................................................................................................................. 9  
6. Data .......................................................................................................................................................... 10  
7. Raw data .................................................................................................................................................. 10  
8. City safety vs. Non-City safety cars dataset .......................................................................................... 11  
9. Weather and parts sales dataset ............................................................................................................. 12  
10. Decision tree dataset ............................................................................................................................. 14  
11. Methods .................................................................................................................................................. 16  
12. Effect of City Safety on Accidents ......................................................................................................... 16  
13. Effect of Weather on Body & Paint parts sales ....................................................................................... 23  
14. Decision Tree for count data ................................................................................................................. 26  
15. Results ................................................................................................................................................... 28  
16. Effect of city safety technology on accidents: ....................................................................................... 28  
   This indicates that city safety has 33% high accident rate then insurance car models. ......................... 29  
17. Effect of Weather and time variables ..................................................................................................... 30  
18. Decision Tree for count data ................................................................................................................. 33  
19. Discussion: ............................................................................................................................................ 35  
20. Effect of city safety technology on accidents: ....................................................................................... 35  
21. Effect of Weather and time variables ..................................................................................................... 37  
22. Conclusions ........................................................................................................................................... 39  
23. Literature: ............................................................................................................................................... 40  
24. Appendix ............................................................................................................................................... 42  
25. Effect of city safety technology on accidents (Graphs and Tables)....................................................... 42  
26. Effect of Weather and time variables (Graphs and Tables) ................................................................. 43
1 Introduction

1. Background

Volvo was founded in Sweden in 1927. Today, Volvo cars corporation (VCC) appears to be a relatively small player in terms of the global car industry, with 373,525 cars sold in 2010 (Tannou & Westerman, 2012).

Volvo Body and Paint business plays an important role in the company’s business and highly contributes to sales in VCC. According to Figure 1, Body and Paint business contributes 32.6% of annual sales in Sweden and 32.8% of annual sales in the rest of the world.

Volvo business has expanded throughout the world, with its largest markets for cars in USA, Sweden, China, Germany, UK and Belgium.

The core of Volvo business is the selling of cars and car parts. The main focus of this study, however, lies solely with the latter. The car parts are divided into various groups and subgroups, each of which in its own manner contributes to the business. Each subgroup is categorized according to its function; for example, Body and Paints subgroup has various function groups, such as radiators, headlamps, windscreens, painted bumpers and unpainted bumpers etc. Apart from function groups, the parts also have specific names and numbers, through which they are easily identifiable in the system. The contribution of Body and Paints subgroup to Volvo overall sales worldwide and in Sweden during 2010 and 2011(with contribution of some important function groups) is shown on the pie chart on Figure 1. It is clear from the bar line plot that each function group decreases with varying percentage rate.
One of the most important function groups in Body and Paints group, for example, is “painted bumper” which constitutes, as from Figure 1, roughly 5% of the group sales.

In comparison with 2010, in the year 2011 Volvo cars (insurance & warranty cars) had 6% less accidents, which consequently resulted in 16% decrease in Body and Paints sales in 2011. According to [European accident research and safety report 2013], in 2009 around 35,500 fatalities were registered in traffic accidents on the territory of European countries. This was a 10% decrease compared to the previous year (2008), which can be considered as a positive sign for Volvo city safety cars department, which strives for achieving “crash-free” results by 2020[Toward crash free cars: Volvo car corporation].

Since the Body and Paint branch constitutes a significant part of the company overall sales, it is important for Volvo Car Corporation to identify the possible causes standing behind the decrease in the number of accidents during 2011 and to isolate and quantify the factors which have direct effect on decrease in Volvo body and paint part sales.

![Figure 1](Importance of Body and Paint Business)
**Objective**
The overall objective of the thesis is to identify the important factors affecting Volvo Body and Paints business, as well as to isolate and quantify the effect of each factor on the sales of the parts in Sweden.

The specific aims of thesis are:
1. Isolate important factors affecting Body and Paint business.
2. Quantify the effect of each factor on Body and Paint business.
3. Make a comparison between city safety cars vs other car models.

**Outline of Thesis**
The thesis is divided into several studies.

The first Study is centered on the comparison between city safety cars and non-city safety cars, to analyze and identify if there is any difference in the number of accidents for two types of car model.

The second study focuses on the relationships between Volvo Body and Paint parts sales in Stockholm, air temperature, precipitation and time variables.

The final study is a tree representation for the total number of accidents in 2010 and 2011 used to identify the categorical factors which play an important role in causing the accidents.
2 Problem discussion
It is important to understand Body and Paints business and to analyze the relationship between accidents and external factors behind them (weather, mileage, city safety, etc.). The point to consider is that VCCS body parts sales are directly proportional to the number of accidents, as an accident potentially results in that the car owner contacts Volvo dealer in order to arrange the repair.
3 Data

1. Data sources
The data sources for both monthly and daily sales for VCCS business were obtained from Volvia which is based in Gothenburg, Sweden. Volvia is a Swedish insurance company, which specializes on Volvo cars and has more than 30% of the customers who own a Volvo car.

Alongside Volvo car owners, Volvia provides services to customers owning Renault, Land Rover and Jaguar cars.

Each year Volvia arranges around 400,000 insurance contracts for Volvo cars. The customers are either insured or warranty based.
2. **Raw data**

1. **City safety vs. Non-City safety cars dataset**

In this study, we will use the data from Sweden, during the period January 2010 till December 2011, and compare city safety cars with non-city safety cars (warranty cars and insurance cars).

![Table of data](image)

**Figure 2 (A subset of the monthly dataset)**

The screenshot shown in Figure 6 (city safety technology sensor) contains a part of data in which accidents are considered as response variable, and other factors such as contracts, year and month are predictors that will be considered in the model building process.
2. Weather and parts sales dataset

In order to see the effect of weather and time variables on the overall sales in Sweden, it is relatively difficult to consider the whole data, as there is no linkage between daily weather and the parts sales throughout the whole Sweden. So it was important, instead, to turn attention to a proxy. Figure 3 contains a comparison between Stockholm monthly sales data and sales data of the entire Sweden. From Figure 3 it becomes clear that Stockholm data can be used as a proxy for the whole country, as there is a high degree of similarity in peaks and troughs between the two plots.

![Figure 3 (Total claims vs. Stockholm Claims)](image)

Weather data was taken from Swedish National databases; temperature and precipitation were the common predictors, where the temperature was measured in °C and precipitation in millimeters.

Snow and rain might also have an effect on daily sales which can also be derived from temperature and precipitation.

The following criteria were considered for transformation:
If temperature is less than -1.1 and precipitation is greater than 0 then mark precipitation as snow.

If temperature is greater than -1.1 and temperature is less than 3.3 then

\[
\text{Snow} = \frac{3.3 \times \text{temperature}}{\text{precipitation} \times (3.3 \times \text{temperature})}
\]

\[
\text{Rain} = \text{precipitation} - \text{snow}
\]

Figure 4 (a subset of daily sales dataset)

Figure 4 shows a table screenshot of daily parts sales dataset from Stockholm; various time and lagged weather variables are mentioned in the table as well.
3. Decision tree dataset

The scope of this study covers car accidents occurred in Sweden during the years 2010 and 2011. The data obtained from Volvia databases was relatively unclear; some of the variables had outliers and missing variables. Hence, the data cleaning, which involves checking completeness of data records and missing values and error removing, was performed at first.

The final dataset was organized based on 8 predictors and 1 response variable which is the frequency of accidents by Volvo cars. The final data set contains the total number of 710 records. The description of each predictor is explained in Table 1.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Role</th>
<th>Variable type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident frequency</td>
<td>Target</td>
<td>Count/Categorical</td>
<td>Frequencies of car accidents</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1: &lt;400 (&lt;400)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2: &lt;300 (&lt;300)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3: &lt;200 (&lt;200)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4: &lt;100 (&lt;100)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5: &lt;50 (&lt;50)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6: &lt;10 (&lt;10)</td>
</tr>
<tr>
<td>Gender [Car Owner]</td>
<td>Input</td>
<td>Categorical</td>
<td>Sex of Car owner (Male/female)</td>
</tr>
<tr>
<td>Mileage</td>
<td>Input</td>
<td>Categorical</td>
<td>Car traveled in 1 year</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1: 0-10,000 (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2: 10,000-15,000 (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3: 15,000-20,000 (3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4: 20,000-25,000 (4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5: &gt;25,000 (5)</td>
</tr>
<tr>
<td>Contracts</td>
<td>Input</td>
<td>Continuous</td>
<td>Total number of cars registered in that year</td>
</tr>
<tr>
<td>Car Type</td>
<td>Input</td>
<td>Binary</td>
<td>CC = company car</td>
</tr>
<tr>
<td>Age of car</td>
<td>Input</td>
<td>Continuous</td>
<td>PC = Private car</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt; 6 years old cars were taken (car age = #of days)</td>
</tr>
<tr>
<td>Month</td>
<td>Input</td>
<td>Categorical</td>
<td>Categorical</td>
</tr>
</tbody>
</table>
|                | Input  | Categorical       | AS = 6,7,8,9,10,11  
DM = 12,1,2,3  
AJ = 4, 5 |
|----------------|--------|-------------------|---------------------|
| Weekday        |        |                   | 1-7(Monday, Tuesday,  
Wednesday, Thursday,  
Friday, Saturday, Sunday) |
| Year           |        | Binary            | Year 2010 and 2011  |
4 Methods

1. *Effect of City Safety on Accidents*

As stated in the article (Tannou & Westerman, 2012); the decrease in the number of accidents was expected to be around 5% per year. According to accidents and contract dataset data provided by several insurance companies (in particular, Volvia and IF AB Sweden), we found almost a 6% decrease in 2011 accidents compared to accidents occurred in 2010. The decrease in 2011 is due to some external and internal factors.

In Sweden, winter season is a major factor since Sweden often has cold weather and hence snowy or icy roads, which causes high numbers of accidents. It was also noted that 2010 winter season was much harsher than 2011 as shown in Figure 5.

![Figure 5(Weekly temperature for Stockholm)](image)

There are also some other important factors, e.g. city safety system; this section will focus on the comparison between city safety and non-city safety cars. It is done in order to analyze if there are any differences in the number of accidents for both types of car models.
Even if there is lot of safety systems that were introduced previously by Volvo Corporations, in this thesis however, we will particularly see how city safety works under some circumstances.

City safety technology works very effectively on low speeds, which has resulted in high decrease in accidents right after the technology was introduced on the market. We will thus consider city safety technology as an important factor, for it seems to have high influence on VCCS parts sales.

City Safety technology monitors the traffic in front of the car with the help of a laser sensor that is built into the windscreen’s upper section, as shown in Figure 6. It can detect the rear end of a vehicle in front of the car. If the driver is about to drive into the vehicle in front of him and does not react in time, the car brakes automatically. The scope of the technology is occasional: everyday low-speed scenarios, such as traffic jams or entering roundabouts - the situations where a large portion of collisions occur due to distracted drivers [MartinDistner et al 2008].

The importance of city safety cars is to reduce the amount of low-speed crashes and causation of low-speed collisions, which mainly occur because of distractions or

---

1 Martin Distner et al, 2008, city safety- A system addressing rear-end collisions at low speeds, Volvo Cars Sweden
inattentions of a driver. In the US a sample of 100 cars accidents was analyzed to investigate whether there is a relation between the collisions and inattention or distraction of the drivers involved. According to the report (the first analysis of such kind, where the researchers collected detailed information on a large number of near-crash events), nearly 80% of all analyzed crashes and 65% of all near-crashes involved driver inattention exactly prior to the onset of the conflict (Neale et al. 2005). Analyzing the UK National accident database (STATS19) for the period starting from 2005, Grover et al. (2007) found that in 44% of the situations the drivers took no avoiding action prior to the collision.

The objective of this section is to evaluate the effectiveness of city safety technology in terms of avoided crashes, by using real life crash data. The rate of accidents in the cars equipped with the safety was compared with the corresponding rate for other Volvo car models such as older warranty cars 2010 and insurance car.

As the response variable is “count data” (i.e. Data that consists of count observations), there are a number of limitations related to the application of a technical model for analyzing the effects of city safety technology.

The rate of accidents was estimate by the number of claims frequency per insured vehicle years.

\[ r_{w,CS} = \frac{Accidents_{w,CS}}{Contracts_{w,cs}} \quad \text{.... Eq. 1} \]

Where

\[ Accidents_{w,CS} = \text{Number of accidents with City Safety} \]

\[ Contracts_{w,cs} = \text{Number of Registered cars with City Safety} \]
The total number of accidents was aggregated on month where the sample size is 72 observations. SAS 9.2 were used to build a Poisson and negative binomial regression model.

**Poisson Regression**

Poisson regression analysis is a technique used to model independent data variables that describes count data (Cameron et al, 1998). It is often applied to study the occurrence of small numbers of counts as a function of a set of predictor variables in experimental and observational studies in many disciplines, including Economy, Demography, Psychology, Biology and Medicine (Gardener et al, 1995).

In this case Poisson regression is used to model the relationship between the rate of Accidents and potential useful independent variables (year, month, car type)

The number of accidents over contracts can be considered by using Poisson regression with contracts as an offset value. The 95% CI for the rate was calculated by using normal approximation to this distribution

\[
\text{Rate}_{w,cs} \pm 1.96 \sqrt{\frac{\text{Rate}_{w,cs}}{VY_{w,cs}}}
\]

The Poisson regression often uses the log link function, which ensures that all of the predicted values of the dependent variable will be nonnegative (Montgomery et al, 2006).

In particular, let \( Y_i, i = 1, \ldots, n \) be \( n \) random variables representing the dependent variable and \( x_{i1}, \ldots, x_{ik}, i = 1, \ldots, n \) be the corresponding values of the \( k \) independent variables. For Poisson regression, \( Y_i, i = 1, \ldots, n \) are modeled as independent Poisson variables. The expected value of \( Y_i \) is linked to a linear function of the independent variables, using a log link function:
\[
\ln E(Y_i|x_{i1} \ldots x_{ik}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}
\]

Poisson Regression model with explanatory and offset variable can be represented as

\[
\log (\lambda_i) = \sum_{j=1}^{15} \beta_{i_j} \cdot \text{Month}_{i_j} - \log(\text{Contracts}) + \beta_{c} + \beta_{12} \cdot \text{city safety}_i + \beta_{14} \cdot \text{warranty2010}_i \\
+ \beta_{15} \cdot \text{insurance}_i + \beta_{16} \cdot \text{year2011}_i + \beta_{17} \cdot \text{year2011}_i
\]

Equation 1

Where \(\beta_0 \ldots \beta_k\) are the parameters of the model; In Poisson regression there are categorical variables such as Month, year and Car type. Categorical variables are converted into dummy variables with \(\ell = 12\) levels minus 1 binary dependent variables for variable month.

**Negative binomial Regression:**

There might be some situations that the Poisson regression may not fit the data well and hence the fitted data may contain over dispersion. Over dispersion may arise when the variance of response is greater than its mean. In order to overcome this problem and improve the results the negative binomial regression is used. In negative binomial regression function over dispersion is taken as a parameter which minimizes the heterogeneity in count data. This model is a generalized model of Poisson regression where it predicts \(\mu_i\) on the base of \(X_i\) and dispersion parameter \(\vartheta_i\).

\[
\mu_i = \exp(X_i \beta + \vartheta_i) = \exp(X_i \beta) \ast \exp(\vartheta_i)
\]

Note that \(\vartheta = 0\) then the model will becomes Poisson regression model. Thus the negative binomial regression has higher flexibility then the Poisson regression. Also as the \(\vartheta\) is smaller the negative binomial approaches the Poisson regression model (Colin Cameron & Pravin K. Trivedi, 1998).
Regarding the comparison of the model performance in terms of accident statistics, particularly the XC60 model appeared to be the most desired by Volvo customers. A new reference from the Highway Loss Data Institute (HLDI) (Russ Rader, 2011); indicates that Volvo XC60 midsize SUVs equipped with standard City Safety system has much fewer registered accidents in comparison with vehicles without the safety feature. Table 2 shows accidents rates of insurance car pool and warranty cars pool.

Table 2 (Claims per contract for different car models)

<table>
<thead>
<tr>
<th>Car Models</th>
<th>Claims 2010</th>
<th>Claim/Contract</th>
<th>Index</th>
<th>Claims 2011</th>
<th>Claim/contract</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>XC90(275)</td>
<td>1668</td>
<td>0.22</td>
<td>139</td>
<td>1380</td>
<td>0.19</td>
<td>102</td>
</tr>
<tr>
<td>C30(533)</td>
<td>1498</td>
<td>0.20</td>
<td>124</td>
<td>1548</td>
<td>0.20</td>
<td>110</td>
</tr>
<tr>
<td>Others</td>
<td>35338</td>
<td>0.16</td>
<td>102</td>
<td>32789</td>
<td>0.17</td>
<td>91</td>
</tr>
<tr>
<td>XC70(295)</td>
<td>1829</td>
<td>0.16</td>
<td>98</td>
<td>1348</td>
<td>0.13</td>
<td>71</td>
</tr>
<tr>
<td>S40(644)</td>
<td>1644</td>
<td>0.13</td>
<td>83</td>
<td>1143</td>
<td>0.11</td>
<td>62</td>
</tr>
<tr>
<td>S60 II(134)</td>
<td>108</td>
<td>0.06</td>
<td>40</td>
<td>923</td>
<td>0.18</td>
<td>98</td>
</tr>
<tr>
<td>V60(155)</td>
<td>28</td>
<td>0.04</td>
<td>25</td>
<td>1682</td>
<td>0.13</td>
<td>72</td>
</tr>
<tr>
<td>XC60(156)</td>
<td>1314</td>
<td>0.16</td>
<td>100</td>
<td>2546</td>
<td>0.18</td>
<td>106</td>
</tr>
<tr>
<td>Total</td>
<td>71978</td>
<td>0.17</td>
<td></td>
<td>67414</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>
Consider Figure 7 which graphs the accident rate for city safety cars and new warranty cars (model year 2010). The graph shows that the accident rates for two types of cars are very alike, which suggest that the safety characteristics of the new car models are also improving, along with the technological developments in the field. Older warranty cars has higher accident rate.

![Model year 2010](image)

Figure 7(Accident rate for different group cars)
2. **Effect of Weather on Body & Paint parts sales**

Weather is an important factor that directly affects accident statistics, and thus is related to changes of VCCS body and paint branch sales dynamics. During winter season sales of body and paint parts in Sweden increase; during summer season, conversely, much lower amount of parts is regularly expected to be sold. However, another aspect, which directly contributes to notably higher sales during winter season, is the more serious damages cars are possibly exposed to in harsher winter conditions, which causes higher probability of a more excessive demand for expensive parts. Following the above, we can presuppose that any increase in the number of accidents can possibly lead to an increase in Volvo’s Body and Paint branch sales.

According to several studies, snow appears to be a central cause for traffic chaos (see e.g. Thornes, 2005; London Assembly, 2009). The impact also varies considerably from study to study; Smith (1982), for example, encountered only an increase of accidents by just 2.2%, whereas the other studies have reported almost double increase in the accident rate (Codling, 1974; Andreescu and Frost, 1998; Suggett, 1999).

The tests show that road surface reaches its most slippery condition when the temperature is close to zero degrees Celsius (Moore, 1975). Campbell (1986), however, researching the same topic in Winnipeg, Canada, found that the number of accidents within the temperature range below -15°C was surprisingly higher than within the temperature range -15°C to 0°C. Apparently not only snow, ice, rainfall, wind, fog or low sun are the factors that contribute to traffic accidents – even hot temperatures (>34°C) have shown to be a contributing factor in Saudi Arabia (Nofal and Saeed, 1997). Other factors can also affect driving; for instance, sudden illnesses (Lam and Lam, 2005) or drink driving (Meyhew et al., 1986; Horwood and Fergusson, 2000; Evans, 2004). Even superstition can play a significant role in causing an accident; Nähä (2002) has discovered that there was an
increase in the amount of fatal accidents on Friday 13\textsuperscript{th} (compared with other Fridays), by 1.63 for women and 1.02 for men accordingly. Fatal accidents among female drivers occurred most often in the temperature interval -3°C to 1°C, which coincides with the slippery road conditions.

In addition, the reports show that a lot of other weather and time factors are quite important as well, as harsh snowy March 2010 was said to have brought more accidents than an average mild winter month. Cool, rainy spring makes people more careful with regard to driving. During summer season, on the other hand, for short distances most people prefer to use bicycles (going to a nearby mall, for example). Moreover, the fact that sometimes people do not initiate the repair immediately after an accident results in the occurrence of lagged time effects.

All of the above mentioned aspects are the important factors which have to be considered within the framework of such research. Using subset algorithms, we can choose the most appropriate factors for a particular model under the analysis.

Selecting the best subset model requires us to search for all possible subsets; e.g. if we have \((p-1)\) predictors, then the best subset algorithm constructs almost \(2^{p-1}\) alternative models.

Model selection procedures, also known as subset selection or variable selection procedures, are executed in accordance with particular criteria, which allow identifying the most appropriate model. The criteria for selecting the appropriate parameter in the model is \(R_p^2\) and \(R_{a,p}^2\), and can be defined as

\[
R_p^2 = 1 - \frac{SSE_p}{SSTO}
\]

And \(R_{a,p}^2\) can be defined
For large values of $n$ or for a multivariate dependent variable the method of generating all subsets becomes infeasible. Thus, polynomial stepwise (greedy) procedures have been proposed. These procedures are based on adding or deleting variables one at a time according to a specific criterion (R. Draper, 1966; R. Hocking, 1976; A. F. Seber, Sen and M. Srivastava). The Forward Selection procedure starts with no variable in the model and adds one variable at a time until either a stopping criterion is satisfied, or all variables are selected. During each step, the variable with the largest single degree of freedom $F$-value among those eligible is considered for inclusion. That is, a variable is added to a $p$-factor regression equation if

$$F_i = \max_i \left( \frac{(\text{RSS}_p - \text{RSS}_{p+i})/\hat{\sigma}_{p+i}^2}{\text{SSE}_p} \right) > F_{\text{IN}}.$$ 

where $(p + i)$ denotes the quantities computed when variable $i$ is added to the current $p$-factors equation. The stopping criterion for the procedure is given by the specification of the quantity $F_{\text{IN}}$.

It is important to see the effect of weather and time variable on daily sales. For this an appropriate model must be selected. The linear regression has the property that it linearly fits the dataset and for this it is important that the variables have a linear relationship with the response variable. A linear model can be stated as

$$E(Y \mid X) = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Where $\alpha$ is intercept and $\beta_j$ are linear coefficient in multi linear regression for $X$ variables. In this models lagged weather variables may have certain effect on Body and Paint part sales in Stockholm and they can be selected by forward selected method.
3. **Decision Tree for count data**

A decision tree is a graphical way to divide the large amount of data into smaller groups or rules and make a decision on response variable. The decision is made by taking predictors as an input variable and target as response variable.

A decision tree model consists of a set of rules for dividing a large collection of observations into smaller homogeneous group with respect to a particular target variable (Yap Bee Wah et al, 2012). It is better that the target variable is categorical variable as the decision tree calculates the probability that a given record belongs to each of target category. Given a target variable and a set of explanatory variables, decision algorithms automatically determine which variables are most important, and subsequently sort the observations into the correct output category (Olson and Yong, 2006).

The common decision tree algorithms in data mining software are CHAID (Chi-Square Automatic Interaction Detector), CART (Classification and Regression tree) and C5. The splitting criteria for CART, C5 and CHAID are Gini index in, entropy and chi-square test respectively (Yap Bee Wah et al, 2012).

The objective of this study is to model the number of accidents over one year time period. For this purpose a classification tree model is developed. The principal behind the CART tree model is to minimize the impurity in the terminal nodes. For this the tree growing method is used which recursively partition the target variable to minimize the impurity in the terminal nodes.

The impurity can be calculated for a node is defined as follows

\[ i(t) = \emptyset(p(1|t), p(2|t), \ldots, p(j|t)) \]  

\[ \ldots \text{Eq. 2} \]
Where $i(t)$ is the measure of impurity of node $t$, $p(j|t)$ is the node proportions and $\emptyset$ is non negative function.  

The measure of node impurity by Gini index in criteria can is defined as

$$i(t) = \sum_{i \neq j} p(i|t), p(j|t)$$  

.... Eq. 2

The partitioning is done by searching all possible threshold values for all input variables to find the threshold that leads to greatest improvement in the impurity score of the resultant nodes. The pruning step will be done to create a sequence of similar trees, through cutting off increasingly important nodes. This step needs complexity parameter which can be calculated through a cost function of the misclassification of data and size of the tree. The last step is to select a tree with right size from prune tree. Large trees normally results in higher misclassification when applied to analyze new data sets. Assessment of the tree is also a valuable step which normally takes pruned tree and test sample as an input parameter.

If the misclassification rate for the test tree is low, the pruned tree will produce a tree like structure diagram and the decision rules whereby important information can be extracted.

---

2 Li-Yen Chang and Wen-Chieh Chen, (2005): “Data mining of tree-based models to analyze freeway accident frequency”

5 Results

1. **Effect of city safety technology on accidents:**
The range of Volvo car accidents frequencies for the sample is from 0 to 72, with mean = 607 and standard deviation = 321. Running Poisson regression the deviance/Degree of freedom is equal to 30 which are far greater than 1 and it indicates over dispersion problem. The deviance of Negative binomial regression is = 1.08 which is much closer to 1 but it makes month variable as insignificant. Thus, from the log likelihood ratio test statistics as shown in Effect of city safety technology on accidents (Graphs and Tables)

Table 7 in Appendix at least two independent variables are significant predictors of frequency of accidents. In Table 3 the value of AIC and BIC for negative binomial regression model is lower than Poisson regression model. Hence the negative binomial regression model demonstrates a better fit then Poisson Regression model.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Poisson Regression</th>
<th>Negative Binomial Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC (smaller is better)</td>
<td>2355.6120</td>
<td>904.9228</td>
</tr>
<tr>
<td>AICC (smaller is better)</td>
<td>2364.1834</td>
<td>905.8319</td>
</tr>
<tr>
<td>BIC (smaller is better)</td>
<td>2389.7620</td>
<td>916.3061</td>
</tr>
</tbody>
</table>

Hence the estimated negative binomial regression model can be written as.

\[
\log(\mu) = \log(Contracts) + 2.24 - 0.57 \cdot citysafety + (0) \cdot warranty2010 \\
- 0.39 \cdot Insurance - 0.29 \cdot year2010 + (0) \cdot year2011
\]

Where \( \mu \) is the number of accidents in month
The estimated Negative binomial regression coefficient, given the other variables are held constant in the model, shows that the difference in the logs of expected counts is expected to be 0.42 units lower then for insurance cars and 0.71 units lower for city safety cars compared to warranty cars.

In order to compare rate of accidents between city safety and Insurance cars models i.e. let $C =$ city safety and $I =$ insurance car models then

$$\log(\lambda_{i=C,ar}) - \log(\lambda_{i=I,ar}) = \log(Contracts_{i=C,ar}) - \log(Contracts_{i=I,ar}) + \beta_{1,i=C} - \beta_{1,i=I}$$

$$\beta_{1,i=I} - \beta_{1,i=C} = \log(\lambda_{i=I,ar}/Contracts_{i=I,ar}) - \log(\lambda_{i=I,ar}/Contracts_{i=I,ar})$$

As from Equation 1 $r$ is the rate of accidents; then $\exp$ (difference in $\beta$) values will give us the ratio between rates of two car models.

$$\beta_{1,i=I} - \beta_{1,i=C} = \log(r_{i=I,ar}/r_{i=C,ar})$$

$$\Rightarrow \exp(-0.42 + 0.71 ) = 1.33$$

This indicates that city safety has 33% high accident rate then insurance car models.
2. **Effect of Weather and time variables**

To see the effect of weather and time variables on sales which are linear regression model. From the output we notice that 55% of variation is explained by the model. All the five variables seem to be significantly important in terms of their effect on the overall sales. Figure shows the whole process of multi linear regression i.e. the data is partitioned into training test and validation datasets. After data partitioning important variables are chosen from the model which has high correlation with log_partsales. Figure shows that partitioning of data is done using stratified partitioning method. Two important variables year and month are used for data partition partitioning. These variables are ordinal and categorical in nature.

![Figure 8(SAS miner window for multiple linear regression)](image)
Figure 9 (sample stratification)

Figure shows training, test and validation error in data set. This also shows that the data fits well with minimum squared error.

Figure 10 (stratified variable list)

Figure 11 (Standard error in training, test and validation)
The $r^2$ square value shows that model has explained almost 55% variations.

From the type 3 analysis of effects almost all variables are significant in the model.

From the estimate values of important variables which are selected by selection model; lagprr (4 days) lagged temperature (2 day) has negative relation with body and paint part sales in Stockholm. Year variable is also an important variable and against year2011 sales data for Stockholm is negative for 2009 and 2010. Accidents variable has highest effect on daily sales.

The effect of each variable in the list is shown as follows

| Variable | Effect | DF | Estimate | Error | t Value | Pr > |t| |
|----------|--------|----|----------|-------|---------|------|---|
| Accidents| Accidents | 1  | 0.0464   | 0.00318 | 14.56   | <.0001 |
| lagPrr4  | LagPrr4 | 1  | -0.00625 | 0.00240 | -2.60   | 0.0099 |
| lagTem1  | LagTemp1| 1  | 0.0109   | 0.00528 | 2.07    | 0.0395 |
| lagTem2  | LagTemp2| 1  | -0.0160  | 0.00539 | -2.96   | 0.0034 |
| Year     | Year2009| 1  | -0.00621 | 0.0193  | -0.32   | 0.7485 |
|          | Year2010| 1  | -0.0553  | 0.0205  | -2.69   | 0.0076 |
3. **Decision Tree for count data**

Two samples were taken from data i.e. our learning or training sample contains 164 observations and test sample contains 124 observations. Before running the `rpart` function in R programming it is important to split the tree according to some criteria. A minimum split of 10 observations was set to create the tree; which means that 10 similar observations will form a split.

In validation if we have >1 similar observation the node code be validated else it would be rejected.

The tree is formed by Gini index method which partitions the tree by searching all possible threshold values for input variables. The partitioning is done by considering greatest improvement in the impurity score of the resultant nodes. When tree is formed from learning dataset then it is important to prune the tree based on critical parameter. A critical parameter has a value which split the tree such as it has less number of nodes and minimum standard error. It is always better to split the tree with less number of splits and minimum error in tree.

<table>
<thead>
<tr>
<th>CP</th>
<th>nsplit</th>
<th>rel error</th>
<th>xerror</th>
<th>xstd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.29167</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.07813</td>
<td>2</td>
<td>0.4167</td>
<td>0.47917</td>
</tr>
<tr>
<td>3</td>
<td>0.04167</td>
<td>4</td>
<td>0.26042</td>
<td>0.39583</td>
</tr>
<tr>
<td>4</td>
<td>0.02083</td>
<td>5</td>
<td>0.21875</td>
<td>0.38542</td>
</tr>
<tr>
<td>5</td>
<td>0.01736</td>
<td>6</td>
<td>0.19792</td>
<td>0.44792</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>9</td>
<td>0.14583</td>
<td>0.48958</td>
</tr>
</tbody>
</table>

```r
modeltree.e = prune(modeltree, cp=.03)
plot(modeltree.e) Text(modeltree.e)
```
Figure 12 (Decision tree for categorical response)

Figure 12 shows that how tree is formed based on minimum split and minimum standard error.

Next step is to make an assessment of the tree based on test sample; the major purpose of this step is to see how much standard error is there if we provide new sample data.

<table>
<thead>
<tr>
<th>Table 5 (standard error calculated at each node)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;10</td>
</tr>
<tr>
<td>&lt;10</td>
</tr>
<tr>
<td>&lt;100</td>
</tr>
<tr>
<td>&lt;200</td>
</tr>
<tr>
<td>&lt;300</td>
</tr>
<tr>
<td>&lt;50</td>
</tr>
</tbody>
</table>

print(err) = 0.1854839

Above calculation shows that we have almost 98% correct tree
6 Discussion:

1. **Effect of city safety technology on accidents:**
The rate of accidents is greatly reduced in cars equipped with city safety compared with other Volvo car models without city safety technology. There are 6 city safety equipped car models in accidents database for year 2010 and 2011. Three cars were considered as insignificant due to small sample size (SSS) and were removed from analysis. Table 6 shows accident rate for different city safety car models.

<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volvo S80 II</strong></td>
<td>SSS</td>
<td>9%</td>
</tr>
<tr>
<td><strong>Volvo V70 III</strong></td>
<td>SSS</td>
<td>12%</td>
</tr>
<tr>
<td><strong>Volvo VC70 III</strong></td>
<td>SSS</td>
<td>8%</td>
</tr>
<tr>
<td><strong>Volvo S60</strong></td>
<td>7%</td>
<td>18%</td>
</tr>
<tr>
<td><strong>Volvo V60</strong></td>
<td>4%</td>
<td>13%</td>
</tr>
<tr>
<td><strong>Volvo XC60</strong></td>
<td>16%</td>
<td>18%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>14%</td>
<td>16%</td>
</tr>
</tbody>
</table>

The comparative analysis was done on rest of Volvo cars equipped with city safety technology. The effect comparison was done by considering all new cars i.e. warranty cars and insurance cars.

The study covers two calendar years in order to cover both summer and winter conditions equally. In Sweden, winters are often means cold weather and hence snowy or icy roads. In order to see the effect of weather and time variables it was difficult to mention the weather variables in the data set as the number of daily accidents will decrease when a particular city such as Stockholm is considered. To overcome this problem month variable is considered as an input parameter which will isolate the weather from the model.
From empirical analysis as mentioned in Table 6 above the rate of accident is below 16% for 2011 year. However for the warranty and insurance cars it was 20% and 10% respectively. This also shows that there is an effect of this technology on number of accidents.

There is also lot of reports published based on finding the effect of city safety technology i.e. [Martin Distner et al 2008] work was based on observational analysis due to lack of real accident data. The report measured that City Safety has potential to reduce the risk of soft-tissue neck injuries in the rear-end impacted car by approximately 60%. The other report was based on real data and was published in 2012 by Volvia AB Gothenburg. In this report only car with city safety technology XC60 was compared with other warranty car models. The report says that reduction in number of accidents decreases by 33%. The only difference in the study in hand and the previous work is that I took whole pool of city safety technology cars and compared with other warranty and insurance cars. While the previous work either only on city safety or was based on observational analysis. In this study Poisson regression model and Negative binomial regression Model is used as tool for comparative analysis. From Poisson regression the effect was calculated as 42% however from Negative binomial regression function the effect was increased to 50% reduction in accidents. However the month variable becomes insignificant in case of Negative binomial regression model.
2. **Effect of Weather and time variables**

If each selected variable has a linear relationship with the daily sales then the appropriate model is multi linear regression. From Figure 13 it is clear that almost all important variables have linear relationship. Some variables have nonlinear relationship due to less number of data points. As it can be seen from plot between rain and sales some of data points are less then 0mm which is technically impossible so the negative rain variables are removed from the model and then fitting linear model to find the effect.

From Figure 13, the relationship between accidents and log partsales suggest that when accidents and contracts gets higher the log of body and paint business in Stockholm also gets higher, this means that both variables has positive and linear relationship with the response.

It is also important that the response should be distributed normally; Figure 14 shows that response variable is normally distributed except for few data points which seem to be outlier or extreme value.
The daily part sales in SEK has a positive skewness coefficient, viz, 1.66 but after converting the daily part sales into log part sales the skewness becomes negative i.e. $= -0.57$. This is also shown using normal probability plots of the daily sales in Figure 14 and 16 in Appendix. As there is skewness and kurtosis in response data, it is considered important to convert it before creating a model.

The actf() in R is used to find the autocorrelation plot. It describes the strength of relationship between different points in the series. In Appendix Figure 17 is used to show such plots for sales in Stockholm on daily basis.

It is also important to see the correlation between each variable; in this instance we use correlation matrix between response and other important variables. Figure 17 shows correlation matrix between all important variables. Correlation between lagged temperature and precipitation variables is decreasing after 3 lagged days. The correlation between accidents and sales are very high and it shows that as accidents increases sales also increases. Temperature and precipitation has negative effect on sales while snow has positive correlation.
7 Conclusions
We have analyzed daily crash data for Stockholm, collected during three year period i.e. from year 2009 till 2011. The major purpose of the analysis was to find and quantify the factors that have a direct effect on the Body and paint business for VCCS.

The study outcomes show that the cars equipped with the city safety system have lower accident rates than those of the insurance cars and other warranty cars. The findings propose that the warranty cars have 50% and 32% high accident rate then city safety and 5 year old insurance cars.

Weather has also a significant impact on road safety. In terms of crash frequency, rate, and severity, winter weather appears to be far more dangerous than wet weather. Most of the weather-related crashes occur during the winter season – during snowfall and when the overall temperature is below -15 °C and the precipitation is higher than 1. The analyses suggest that temperature has an effect of 1.6% on accidents sales. Rain has almost negative effect on sales which means that people drive less or slowly during rain. In addition, the results of the study suggest that the time variables, such as weekday, month and week, also play an important role in the business.

Then, in the tree based approach two models have been implemented. The model with contract variable has shown that increase in the number of cars for a particular model can possibly result in the consequent increase in the number of accidents.
8 Literature:


M Distner, M Bengtsson, T Broberg, L Jakobsson. City safety: Volvo Cars Sweden, paper Number 09-0371


Russ Rader (2011). High Tech System on Volvos is Preventing Crashes.


Neale VL, Dingus TA, Klauer SG, Sudweeks J, Goodman M. An Overview of the 100 Car Naturalistic Study and Findings, Paper No. 05-0400, 19th Int. ESV Conf., 2005


9 Appendix

1. **Effect of city safety technology on accidents (Graphs and Tables)**

   Table 7 (likelihood ratio test statistics for categorical variables)
   
<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>1</td>
<td>22.06</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>type</td>
<td>2</td>
<td>64.32</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

   Table 8 (output from negative binomial regression)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Effect</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>2.2310</td>
<td>.0605</td>
<td>2.1124</td>
<td>2.3495</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>year 2010</td>
<td>1</td>
<td>-0.2925</td>
<td>.0580</td>
<td>-0.4063</td>
<td>-0.1788</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>year 2011</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
</tr>
<tr>
<td>type Insurance</td>
<td>1</td>
<td>-0.4204</td>
<td>-34%</td>
<td>0.0696</td>
<td>-0.5568</td>
<td>-0.2840</td>
</tr>
<tr>
<td>type citysafety</td>
<td>1</td>
<td>-0.7109</td>
<td>-50%</td>
<td>0.0713</td>
<td>-0.8506</td>
<td>-0.5711</td>
</tr>
<tr>
<td>type warranty</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
</tr>
<tr>
<td>Dispersion</td>
<td>1</td>
<td>0.0561</td>
<td>0.0099</td>
<td>0.0397</td>
<td>0.0792</td>
<td>.</td>
</tr>
</tbody>
</table>
2. **Effect of Weather and time variables (Graphs and Tables)**

![Probability Plot of log_partsales](image)

**Figure 14**(Normality plot for response variables)

![Histogram of partsales](image)

**Figure 15**(histogram partsales)

![Histogram of log_partsales](image)

**Figure 16**(histogram for log partsales)
Figure 17 (distribution of log of part sales)

Series `carData$log_partsales`

Figure 18 (residuals vs ordered row count)
Figure 19 (predicted vs original)

log_partsales

Correlations: log_partsales, accidents, Temperature, lagTem1, lagTem2, lagTem3, lagTem4

<table>
<thead>
<tr>
<th></th>
<th>log_partsales</th>
<th>accidents</th>
<th>Temperature</th>
<th>lagTem1</th>
<th>lagTem2</th>
<th>lagTem3</th>
<th>lagTem4</th>
</tr>
</thead>
<tbody>
<tr>
<td>accidents</td>
<td>0.701</td>
<td>0.000</td>
<td>-0.225</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.225</td>
<td>-0.255</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>lagTem1</td>
<td>-0.215</td>
<td>-0.276</td>
<td>0.952</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>lagTem2</td>
<td>-0.210</td>
<td>-0.248</td>
<td>0.818</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>lagTem3</td>
<td>-0.186</td>
<td>-0.235</td>
<td>0.878</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>lagTem4</td>
<td>-0.193</td>
<td>-0.243</td>
<td>0.898</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>snow</td>
<td>0.123</td>
<td>0.145</td>
<td>-0.370</td>
<td>-0.353</td>
<td>-0.347</td>
<td>-0.349</td>
<td>0.343</td>
</tr>
<tr>
<td>rain</td>
<td>-0.109</td>
<td>-0.131</td>
<td>0.340</td>
<td>0.349</td>
<td>0.343</td>
<td>0.343</td>
<td>0.343</td>
</tr>
</tbody>
</table>

Figure 20 (Correlation matrix)
This thesis focuses on identifying the degree of contribution of the most important factors affecting Body and Paint business of Volvo Car Corporation in Sweden. It is clear that Body and Paint business for VCCS directly depends on the number of registered accidents. Our major purpose is to determine the factors which have direct or indirect effect on reduction in the number of accidents in Sweden and to analyze in which degree they may affect the business. During the interviews with senior staff members, we discover that particularly city safety cars are mentioned by most of the specialists. Other important factors highlighted were mileage, weather, company car/private car and age of a car.

City Safety is a technology designed to help the driver mitigate, and in certain situations avoid, collisions at low speed by automatically bracking the vehicle. The estimated claim rate frequency i.e. claims per contract rate was 50% lower for city safety equipped; then other warranty cars models without system. The study also analysis the effect of rain, mean temperature and snow on Volvo Body part sales in Stockholm Sweden. Temperature snow impacted road accidents significantly. Snow was shown to be the leading variable, as the number of accidents increases sharply with increased snowfall. Temperature is the second important variable in the list i.e. as the temperature decreases by 1° C the sales of body and paint business in Stockholm increases by 1.6%.

Time variable such as weekday, month, and year also plays significant role in this model. During Fridays 51% high accidents are expected then accidents occurred on Sundays.

Keywords
Volvo Car customer service, City safety technology, accident rate, machine learning models, Poisson Regression, Negative binomial regression, CART