An Empirical Analysis of Stockholm Bilpool Using Generalized Linear Model

Author: Yang Li, Xiu Xie
Supervisor: Mikael Möller

June 10, 2009

D-level Essay in Statistics in Spring 2009
Department of Economics and Society, Dalarna University
An Empirical Analysis of Stockholm Bilpool Using Generalized Linear Model

By Yang Li, Xiu Xie

June 10, 2009

Contents

1 Introduction 3
  1.1 Background of GLMs .................................................. 3
  1.2 Background of Carsharing ............................................ 3
    1.2.1 History of Carsharing ........................................... 3
    1.2.2 Definition ......................................................... 4
    1.2.3 How It Works ..................................................... 4
  1.3 Aim of the Essay ....................................................... 4

2 Data Description 4
  2.1 Original Data .......................................................... 4
  2.2 Processing Data ....................................................... 5
    2.2.1 Addition .......................................................... 5
    2.2.2 Deletion .......................................................... 5
    2.2.3 Partition .......................................................... 5
    2.2.4 Selection ........................................................ 6

3 Methodology 7
  3.1 Conception of GLMs .................................................. 7
  3.2 Definition of GLMs ................................................... 7
  3.3 Properties of GLMs .................................................. 8

4 Model and Analysis 9
  4.1 The Response Distribution ........................................... 9
  4.2 Results and Discusses ............................................... 11

5 Conclusion 13
  5.1 Parameter Explanation ............................................... 13
  5.2 Suggestions ......................................................... 15
Abstract

With a further acquaintance of Carsharing, more and more people are showing their great interest into a new trend of sharing cars which has become a new substitution of the present transportations including owning a car or renting a car. Carsharing is a new transportation pattern which not only has the economical characteristics of resources sharing of public transport and also protects the people privacy as private cars do. This essay applies generalized linear model(GLMs) to conduct an empirical analysis of Stockholm Bilpool and is aiming to find some factors that affect the income of the Carsharing Organizations and to suggest some possible solutions to rise the income. We find that factors such as year, month, size of the car, city or suburb, workdays or holidays and the distance between member’s home and the Bilpool locations affects the income. We suggest that the CSOs maybe can rise the income by changing the price at different time in a year and showing more care to the people who live farther from a Bilpool. Also we advise that CSOs can adjust prices of the cars of different size in different time of a year.

Key words: Carsharing, GLMs, Year, Month, City, Income, Distance, Size

Not everyone can have a car if we still want a planet - unless we change.¹

¹LEE SCHIPPER
1 Introduction

1.1 Background of GLMs

Generalized linear models (GLMs) is one of the most common used models in modern statistic analysis nowadays. As most statisticians know, logistic regression and probit regression are techniques commonly used for modeling a binary response variable as a function of one or more predictors. These techniques have a long history, with the term "probit" traced by David (1995) [1] back to Bliss (1934) [2], and Finney (1952) [3] attributing the origin of the technique itself to psychologists in the late 1800s. GLMs appeared in the statistical scene in the article of Nelder and Wedderburn (1972) [4]. They are the first ones to put the unified regression methodology of the former works showing the similarities between seemingly disparate methods. With the development of computer science, the application of GLMs becomes a reality in the mid 1970s. And now GLMs has become a mature data-analytic methodology (Lindsey 1996 [5]) and have been developed in numerous directions. There are techniques for choosing link functions and diagnosing link failures (Mallick and Gelfand 1994 [6]; Pregibon 1980 [7]) as well as research on the consequences of link misspecification (Weisberg and Welsh 1994 [8]). There are techniques for outliers detection and assessment of case influence for model checking (Cook and Croos-Dabrera 1998 [9]) and methods of modeling the dispersion parameters as a function of covariates (Efron 1986 [10]) and for accommodation measurement error in the covariates (Buzas and Stefanski 1993 [11]; Stefanski and Carroll 1990 [12]) as well as ways to handle generalized additive models (Hastie and Tibshirani 1990 [13]). There are approaches pioneered known as generalized estimating equations (GEEs) (Liang and Zeger 1986 [14]) contribute an accommodation of a wide array of correlated data structures and the popularization of the "robust sandwich estimator" of the variance-covariance structure.² [15]

1.2 Background of Carsharing

Nowadays, car has become one of the major tools in the transportation in human’s life. Possessing a car equals to faster speed, comfortable condition and privacy which are all great benefits. But the fast expansion of the car using is negatively affecting the living condition of people. Excessive cars cause traffic congestion, air pollution, global warming, noise problem, parking problem, oil waste and etc. People are aware of the problems and come up with many ideas to solve them. Sharing cars is an effective and economical solution.

1.2.1 History of Carsharing

The first reference to carsharing in print is the Selbstfahrergenossenschaft carshare program in a housing cooperative that got underway in Zurich in 1948, but there was no known formal development of the concept in the next few years. The early 1970s saw the first whole-system carshare projects. The 1980s and first half of the 1990s was a "coming of age" period for carsharing, with continued slow growth, mainly of smaller non-profit systems, many in Switzerland and Germany, but also on a smaller scale in Sweden, the Netherlands, Canada and the United States. This 1990s wave of innovation continued into

²Generalized linear models; Charles E. McCulloch
the present decade, with carshare developments advancing at different speeds in different countries, but with a generally accelerating pace when taken in sum.\(^3\)

In May 2002 the Swedish National Road Administration published the report "Bilpooler - nyckeln till flexibelt resande" ("Car-sharing - the key to combined mobility"). Much has happened since then, the development of car-sharing in Sweden having been accelerated.

1.2.2 Definition

Car-sharing means that a number of persons share the use of one or more cars. Use of a car is booked beforehand, the user paying a fee based on the distance driven and the length of time the car was made use of.\(^4\)

Car-sharing is a different concept from the traditional car rental.

- Users must be members of the car-sharing company (mainly residents) and pre-proved to drive.
- There is no limited office time and vehicles can be rented by hours or by days.
- It’s all self-service (reservation, pickup and return) and no after service (clean, petrol filled up).
- Vehicle locations are distributed throughout the service area, and often located for access by public transport.
- Insurance and fuel costs are included in the rates.

1.2.3 How It Works

The CSOs (Car Sharing Organisations) require a checking of past driving records and an annual fee from the users. The vehicle is reserved in advance, usually over the Internet or telephone (and increasingly by mobile phones). Most companies charge an hourly fee for the time that the car is in use plus a fee per km driven. If a vehicle is not returned at the scheduled time, a high penalty will be charged, for it may interfere with other drivers’ reservations. Members are responsible for leaving the vehicles on time, in the agreed parking area, clean and in good condition for the next user.

1.3 Aim of the Essay

In this article we will fit a Generalized linear model to analyse the factors that affect the income of the Stockholm Bilpool in 2007 and 2008. With this understanding, we try to suggest some possible solutions to increase the income.

2 Data Description

2.1 Original Data

We obtained the data from Stockholm Bilpool. The original data contains details of each rental (a whole event of a member booking a car, picking up a car, using a car and returning the car) and information

\(^3\)wikipedia
\(^4\)the Swedish National Road Administration
Examples of the original data set is given in table 1.

Table 1

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>Bilpool Location(BLoc)</th>
<th>Car</th>
<th>Plate</th>
<th>UserID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>1</td>
<td>Fridhemsplan</td>
<td>Focus</td>
<td>WKC216</td>
<td>91</td>
</tr>
<tr>
<td>2007</td>
<td>2</td>
<td>Bagartorp</td>
<td>Avensis</td>
<td>TTS674</td>
<td>233</td>
</tr>
<tr>
<td>2008</td>
<td>4</td>
<td>Medborgarplatsen</td>
<td>Aygo</td>
<td>GKF857</td>
<td>232</td>
</tr>
<tr>
<td>2008</td>
<td>5</td>
<td>Rosenlunds</td>
<td>Fiesta</td>
<td>XNC822</td>
<td>209</td>
</tr>
</tbody>
</table>

... ...

Follow Table 1

<table>
<thead>
<tr>
<th>Hourfee</th>
<th>Distancefee</th>
<th>bookingfee</th>
<th>Trip-start</th>
<th>Trip-stop</th>
<th>Mins</th>
<th>KM</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>1.75</td>
<td>15</td>
<td>2007-01-13</td>
<td>12:52</td>
<td>242</td>
<td>39</td>
</tr>
<tr>
<td>17</td>
<td>1.95</td>
<td>15</td>
<td>2007-02-03</td>
<td>10:12</td>
<td>151</td>
<td>19</td>
</tr>
<tr>
<td>17.5</td>
<td>1.6</td>
<td>15</td>
<td>2008-04-07</td>
<td>17:11</td>
<td>832</td>
<td>72</td>
</tr>
<tr>
<td>17.5</td>
<td>1.6</td>
<td>15</td>
<td>2008-05-06</td>
<td>11:03</td>
<td>262</td>
<td>69</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

2.2 Processing Data

2.2.1 Addition

According to what we need, we create several new variables and put them into the data set.

- "Income" means the money the CSOs get from a rental. We calculate the income by the formula.

\[
Income = Hourfee \times Mins \div 60 + Distance \times KM + Bookingfee
\]

- "Dis" means the distance between the Bilpool location and the member’s living place. We use the GOOGLE MAP to measure the distance between the Bilpool address and the member’s home address. (We choose walking distance when the distance is under 2 km, otherwise we choose driving distance.)

2.2.2 Deletion

When studying the data, we discovered that there exist some inconsistent and irrelevant observations, so we need to delete them. We delete the observations of which KM column is 0, which suggestes that the member did booked a car but ended up not picking it up for some reason. These observations will interfere the analysis, so we delete them.

2.2.3 Partition

Aiming for a clearer result, we need to group 2 variables which have too many catagories.
"Size" denoted as 0,1 which means the car is medium or small size. There are 8 Brands of cars in the data set. But in these 2 years, there are some replacements among cars. So we decide to group the cars by their size(big, medium, small). But we find that only one Avensis car(big, Plate:TTS674) was rented 78 times in the 2007 and there's no big cars anymore since 2008. Comparing to the other 2 groups, the data of big car is too small to be a group. So we decide to put them into the medium size car group and analyse the medium size and the small size car.

"City" denoted as TRUE or FALSE which means that the Bilpools are in the city or in the suburb. Here is a map of the location of the Stockholm Bilpools that we investigated.

We can see the locations of the 8 observed Bilpools in the Map and we decide to divide the eight locations into two groups. One group is the city group with Fridhemsplan(fn), Gärdet(gt), Medborgarplatsen(mn), and Rosenlunds(rs), and the other group is suburb group with Bagartorp(bp), Bergshamra(ba), Gubbängen(gn), and Solna(sa).

2.2.4 Selection

Finally, we choose some data as the variables from the original data set.

5Aygo was bought in 2007 and replaced Yaris. Corsa has been replaced by Ford Fiesta. Avensis was replaced by Corolla.
60=medium=Focus, Astra, Corolla, 1=small=Corsa, Fiesta, Yaris.
7TRUE: the Bilpools in the city; FALSE: the Bilpools in the suburb.
• "Income" means the money the CSOs get from user in a contract.
• "Year" denoted as 2007 and 2008.
• "Month" denoted as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 (from Jan to Dec).
• "Dis" means the distance between the bilpool location and the member’s living place.
• "WDay" denoted as TRUE, FALSE which means the day when a person rent a car is work day or holiday.\(^8\)
• "Size" denoted as 0,1 which means the car is medium or small size.

3 Methodology

3.1 Conception of GLMs

A generalized linear model is different from a General linear model. There is no error term. Instead, we model the expected value, \( E(y) \), considering \( y \) directly as a random variable. A summary of the conception of GLMs are listed as following:

• It drops the error term and puts the whole randomness in the response, \( y \).
• It drops the normality assumption and allows the response to have any distribution in the exponential family.
• It drops the linearity assumption in favour of a link function, \( g(\cdot) \), which is assumed to transform the non-linear relation into a linear relationship (which is called a linear predictor).
• It drops the assumption of equal variance in favour of a variance function that explains how the individual variation is related to its mean.

3.2 Definition of GLMs

A generalized linear model consists of three components:

• A random component, specifying the conditional distribution of the response variable, \( Y_i \), given the predictors.

Traditionally, the random component is an exponential family — the Normal (Gaussian), Binomial, Poisson, Gamma, or Inverse-Gaussian family of distributions — but generalized linear models have been extended beyond the exponential families.

Exponential families of distributions can be written in the form

\[
f(y_i) = \exp \left( y_i \theta_i - b(\theta_i) \right) a_i(\phi) - c(y_i, \phi)
\]

where \( a_i(\phi), b(\theta_i), \) and \( c(y_i, \phi) \) are known functions, \( \theta_i \) is called the canonical parameter; and \( \phi \) is a dispersion parameter.

\(^8\)TRUE: workdays; FALSE: holidays
In all models considered in these notes the function \( a_i(\phi) \) has the form

\[
a_i(\phi) = \frac{\phi}{p_i}
\]

where \( p_i \) is a known prior weight, usually 1.

- A linear function of the regressors, called the linear predictor

\[
\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}
\]  

(2)

on which the expected value \( \mu_i \) of \( Y_i \) depends. The mean \( \mu_i \) is a function of the canonical parameter \( \theta_i \).

The \( X \)'s may include quantitative predictors, but they may also include transformations of predictors, polynomial terms, contrasts generated from factors, interaction regressors, etc.

An invertible link function

\[
g(\mu_i) = \eta
\]

(3)
transforms the expectation of the response to the linear predictor.\[16\]

### 3.3 Properties of GLMs

Standard link functions and their inverses are shown in the Table 2:

<table>
<thead>
<tr>
<th>Link</th>
<th>( \eta_i = g(\mu_i) )</th>
<th>( \mu_i = g^{-1}(\eta_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity</td>
<td>( \mu_i )</td>
<td>( \eta_i )</td>
</tr>
<tr>
<td>log</td>
<td>( \log e \mu_i )</td>
<td>( e^{\eta_i} )</td>
</tr>
<tr>
<td>inverse</td>
<td>( \mu_i^{-1} )</td>
<td>( \eta_i^{-1} )</td>
</tr>
<tr>
<td>inverse-square</td>
<td>( \mu_i^{-2} )</td>
<td>( \eta_i^{-1/2} )</td>
</tr>
<tr>
<td>logit</td>
<td>( \log e^{\mu_i} )</td>
<td>( \frac{1}{1+e^{-\eta_i}} )</td>
</tr>
<tr>
<td>probit</td>
<td>( \Phi^{-1}(\mu_i) )</td>
<td>( \Phi(\eta_i) )</td>
</tr>
<tr>
<td>complimentary log-log</td>
<td>( \log e[-\log e(1 - \mu_i)] )</td>
<td>( 1 - \exp[-\exp(\eta_i)] )</td>
</tr>
</tbody>
</table>

For distributions in exponential families, the conditional variance of \( Y \) is a function of the mean \( \mu \) together with the dispersion parameter \( \phi \), as shown in the Table 3:

| Family       | Canonical Link | Range of \( \eta_i \) | \( V(Y_i|\eta_i) \) |
|--------------|----------------|------------------------|---------------------|
| Gaussian     | identity       | \(( -\infty, +\infty )\) | \( \phi \)           |
| Binomial     | logit or (probit or log-log) | \( 0, 1, 2, \ldots, n_i \) | \( \frac{\phi \mu_i^2}{n_i} \) |
| Poisson      | log            | \( 0, 1, 2, \ldots, n_i \) | \( \phi \mu_i^2 \) |
| Gamma        | inverse        | \(( 0, \infty )\)       | \( \phi \mu_i^3 \)   |
| Inverse Gaussian | inverse-square | \(( 0, \infty )\)       | \( \phi \mu_i^3 \)   |

Maximum Likelihood Estimation Methods is in the Appendix.
4 Model and Analysis

4.1 The Response Distribution

When we take "Income" as the response of generalized linear model, we should test its distribution first. For this purpose we use histogram and Q-Q plot and start by checking for normal distribution.

Through Figure 1 and Figure 2, we can see that income does not fit the normal distribution well.

When the response does not fit normal distribution, we should transform it to make its distribution normal. In Box-Cox transformation, it is assumed that there is $\lambda \neq 0$ such that a transformation of the observed data $y$ according to

$$y_\lambda = \frac{y^\lambda - 1}{\lambda}$$

has a normal model $N(\mu, \sigma^2)$. The value $\lambda = 0$ is defined to represent the log-transformation. The value $\lambda = -1$ is defined to represent the inverse transformation. The value $\lambda = 1$ is defined to represent no transformation.

The appropriate value of $\lambda$ can be found from the profile log-likelihood of $\lambda$.

$$\log L(\lambda) = -\frac{1}{2} \log \sigma^2(\lambda) - \frac{n}{2} + (\lambda - 1)\Sigma \log y_i$$

In practice we always use simple values of $\lambda$.\[17]

So we construct a linear model with all independent variables, and plot its profile log-likelihoods for the parameter of the Box-Cox power transformation as Figure 3 shows below.
We can see the parameter interval is very short, and near 0. Hence we chose $\lambda = 0$. Then we check whether log(income) belongs to normal distribution by Figure 4. From the normal QQ plot, what can see that the log(income) still not fit the normal distribution well.

But we use log(income) and make a histogram of Figure 5, and it seems like a Gamma distribution. So we check the gamma QQ plot of Figure 6, to see if it’s reasonable to assume that the response is Gamma distribution.
Figure 6 shows that log(Income) fits the Gamma distribution quite well. And the log transformation is easier for us to explain. So we assume that the response distribution is the Gamma distribution. Thus, we take \( y = \log(\text{Income}) \) as response and take Year, Month, City, WDay, Dis and Size as independent variables.

### 4.2 Results and Discusses

We construct a generalized linear model as follows:

- Response: \( E(y) = \mu \)
- Family: Gamma
- Link function: \( g(\mu) = \mu^{-1} = \eta \) (canonical link)
- Linear predictor:

\[
\eta = \alpha + \beta_1 x_{\text{Year}} + \beta_2 x_{\text{Feb}} + \beta_3 x_{\text{Mar}} + \beta_4 x_{\text{Apr}} + \beta_5 x_{\text{May}} + \beta_6 x_{\text{Jun}} \\
+ \beta_7 x_{\text{Jul}} + \beta_8 x_{\text{Aug}} + \beta_9 x_{\text{Sep}} + \beta_{10} x_{\text{Oct}} + \beta_{11} x_{\text{Nov}} + \beta_{12} x_{\text{Dec}} \\
+ \beta_{13} x_{\text{City}} + \beta_{14} x_{\text{WDay}} + \beta_{15} x_{\text{dis}} + \beta_{16} x_{\text{Size}}
\] (4)

<table>
<thead>
<tr>
<th>Definition of variables</th>
<th>( x_{\text{Year}} )</th>
<th>( x_{\text{Feb}} )</th>
<th>( x_{\text{Mar}} )</th>
<th>( x_{\text{Apr}} )</th>
<th>( x_{\text{May}} )</th>
<th>( x_{\text{Jun}} )</th>
<th>( x_{\text{Jul}} )</th>
<th>( x_{\text{Aug}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2007</td>
<td>otherwise</td>
<td>otherwise</td>
<td>otherwise</td>
<td>otherwise</td>
<td>otherwise</td>
<td>otherwise</td>
<td>otherwise</td>
</tr>
<tr>
<td>1</td>
<td>2008</td>
<td>Feb</td>
<td>Mar</td>
<td>Apr</td>
<td>May</td>
<td>Jun</td>
<td>Jul</td>
<td>Aug</td>
</tr>
<tr>
<td>( x_{\text{Sep}} )</td>
<td>( x_{\text{Oct}} )</td>
<td>( x_{\text{Nov}} )</td>
<td>( x_{\text{Dec}} )</td>
<td>( x_{\text{City}} )</td>
<td>( x_{\text{WDay}} )</td>
<td>( x_{\text{Size}} )</td>
<td>( x_{\text{dis}} )</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>otherwise</td>
<td>otherwise</td>
<td>otherwise</td>
<td>Suburb</td>
<td>Workdays</td>
<td>Medium</td>
<td>Small</td>
<td>Numeric</td>
</tr>
<tr>
<td>1</td>
<td>Sep</td>
<td>Oct</td>
<td>Nov</td>
<td>Dec</td>
<td>City</td>
<td>Workdays</td>
<td>Medium</td>
<td>Numeric</td>
</tr>
</tbody>
</table>

We apply iteratively reweighted least squares (IWLS) to fit this model and do the Wald test on the estimated parameters. "t-value" is Wald test statistics calculated under hypothesis

\[
H_0 : \beta_i = 0 \\
H_A : \beta_i \neq 0
\]

and the statistics is

\[
z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}
\] (5)

In this case, \( z_i \) is normal distributed for large sample. The p-value of \( \beta_i \) is the possibility of \( z_i \) in normal distribution. When p-value is less than 0.05, we can reject the hypothesis with confidence level 95%. The coefficients matrix coverges after 4 times of Fisher Scoring iterations and is showed in Table 4.
glm(formula = log(income) ~ Year + Month + City + Dis + WDay + Size, family = Gamma, data = d0)

| Names        | parameters | Estimate | t-value | Pr(>|t|) |
|--------------|------------|----------|---------|----------|
| (Intercept)  | $\alpha$  | +        | 91.756  | ***      |
| Year2008     | $\beta_1$ | +        | 3.616   | ***      |
| Feb          | $\beta_2$ | +        | 1.765   | -        |
| Mar          | $\beta_3$ | -        | -2.157  | *        |
| Apr          | $\beta_4$ | -        | -0.708  |          |
| May          | $\beta_5$ | -        | -2.946  | **       |
| Jun          | $\beta_6$ | -        | -4.196  | ***      |
| Jul          | $\beta_7$ | -        | -8.960  | ***      |
| Aug          | $\beta_8$ | -        | -4.719  | ***      |
| Sep          | $\beta_9$ | -        | -0.178  |          |
| Oct          | $\beta_{10}$ | +      | 0.800   |          |
| Nov          | $\beta_{11}$ | +      | 1.352   |          |
| Dec          | $\beta_{12}$ | +      | 1.122   |          |
| City         | $\beta_{13}$ | -      | -5.484  | ***      |
| Distance     | $\beta_{14}$ | -      | -16.937 | ***      |
| Workdays     | $\beta_{15}$ | +      | 7.504   | ***      |
| Size(small)  | $\beta_{16}$ | +      | 6.044   | ***      |

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

"+" Positive Parameter. "+" Negative Parameter

Figure 7

Figure 8
We can see from Table 4 that the test of some of the estimated parameters are significant in the Wald test. And then we check the residual plot in Figure 7, the residuals fit the normal distribution well and the residual deviance is 115.99 on 4143 degrees of freedom which decrease the Null deviance in some levels of degree.\(^9\)

Figure 8 is an Ellipse Graph that shows the related coefficient of variables. If the ellipse points to the right direction, then it means that the two variables are positive correlated, otherwise, negative correlated. The thinner the ellipse is, the bigger the related coefficient is. From Figure 8, we can see that there isn’t much correlation between each two variables. So we can say the model doesn’t have autocorrelations.

We can see from Table 4 that most P-values are much smaller than 0.05 except for several months. Therefore, we should reject the null hypothesis in a very high confidence level, 95%, and reach to a conclusion that the change of the factors which have pass the test can affect the income. And the intercept includes the influence of Year=2007, Month=January, City=suburb, WDay=holiday, and Size=big. So, the parameters are not the direct influence on income, but the influence of value 1 comparing with value 0.\(^10\) For example, the estimate of Year=2008 means how higher is income of 2008 than 2007, not the influence of Year=2008 on income.

So in which way do these factors affect the income? We give this Table 5 to show the answer.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$x$</th>
<th>$\eta = \mu^{-1}$</th>
<th>$\mu = \eta^{-1}$</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>increase</td>
<td>increase</td>
<td>decrease</td>
<td>decrease</td>
</tr>
<tr>
<td>-</td>
<td>decrease</td>
<td>decrease</td>
<td>increase</td>
<td>increase</td>
</tr>
</tbody>
</table>

5 Conclusion

5.1 Parameter Explanation

**Year** From Table 4 we can see that $\beta_1$ is positive which means that, comparing to 2007, the income is getting less in 2008. We assume that the financial crisis in 2008 is the most likely cause.

**Month** We check the average of "Income" changes with "Month" in Figure 9. We can see that there exists an obvious steady low level of income from September to February and a very intensive fluctuation in a period from March to August.

\(^9\)Model Output in Appendix

\(^{10}\)Refer to the Table "Definition of variables"
We know that $\beta_i$ of March and of May to August are negative and significant in the model, and these 5 months have a higher income than the other months do, consistenting with the results of the model estimation.

We assume some general reasons for this phenomenon.

- **Great peak between May and August.**

Everyone knows that the best weather of the year in Sweden begins from May and lasts to August. And people are encouraged to enjoy a summer trip in many ways, such as vacations. So there's no doubt that people tend to have a longer-time trip in this period of a year.

- **Low level of income from September to February.**

As we know the time from September to April in Sweden is in cold winter and at most of the year it's snowing everywhere. So the weather is the main disadvantage factor for a longer trip. And in September ended the favourite trip time of the whole year when people tend to change their attention to study or work, and also in December people probably are having Christmas at home with families.

- **The small climb-up in March.**

At the end of the winter, people incline to make some short trips to enjoy the last beautiful winter views or to go skating for the last few times before spring comes. And also there is a week of school vacation, so students are probably using cars more often than other time of the year.

**City or Suburb**  From Table 4 we can see that $\beta_{13}$ is negative which means Bilpools in the city tend to make more money than one do in the suburb. We think it's probably because that more members live in the city and that the Hourfee\textsuperscript{11} in the city is higher than that in the suburb.

**Distance**  From Table 4 we can see that $\beta_{14}$ is negative which means that a person lives far from a Bilpool tends to attribute more money to the Bilpool in one time. This is just the opposite to what we have assumed. But this phenomenon did tells us something new that if a person lives far from a Bilpool, then he or she will prefer to make a longer trip than someone who lives closer to the Bilpool. CSOs should pay more attention to this.

**Workdays or Holidays**  From Table 4 we can see that $\beta_{15}$ is positive which means people tend to pay less money on workdays than on holidays on sharing cars. We think it's obvious that people have higher probability of using cars for a longer time in holidays.

**Size(small) or Size(medium)**  From Table 4 we can see that $\beta_{16}$ is positive which means small size cars tend to make less money than the medium size cars. We think it's probably because that people are willing to take a medium size car for a longer trip while a small size car for a short trip.

\textsuperscript{11}Hourfee in city:17.5 sek/hr to 22sek/hr; Hourfee in suburb:11.5kr/hr to 14.5sek/hr
5.2 Suggestions

According to the results above, we give some suggestions about the possible ways to increase the income of the Bilpools.

- CSOs probably can adjust the price in the high income period to encourage people have longer time trip and give more attention to the small climb-up in March. Meanwhile, with regard to the period from September to February, it’s reasonable to assume that it’s hard to encourage people to make longer trip in this period of time, so we consider CSOs can make strategies to promote people to increase the times of short distance car using. For example, up to 20 times per month from September to February, people can get rewards.
- CSOs maybe can change the time fee to make more money.
- CSOs could offer some humanistic perspective services to the people who live farther from the Bilpool to encourage them to come to the Bilpool more times.
- In holidays CSOs should take price adjusting strategies and various promotions to encourage people to have longer trip.
- CSOs maybe can lower the medium car price in the high income period or in holidays to encourage people to have longer trip or can lower the small size car price from the September to February.

What have been written above are some hypothetical suggestions according to the model results and the reasons we assumed. More investigation is needed to prove these propositions.

Appendix

**Maximum Likelihood Estimation** The likelihood function for any distribution of the exponential family can be expressed in matrix notation as

\[
L(y; \theta) = \exp \left( \frac{y^T \theta - 1^T b(\theta)}{a(\phi)} - 1^T c(y, \phi) \right)
\]

(6)

\[
\Rightarrow \log(L(y; \theta)) = l = \frac{y^T \theta - 1^T b(\theta)}{a(\phi)} - 1^T c(y, \phi)
\]

(7)

Though the likelihood function does not explicitly contain the model parameter vector, \( \beta \), it appears in the equation implicitly through \( \theta \). Now, the first differentiation of the log likelihood \( l \) w.r.t \( \beta \) can be written as

\[
Dl(\beta) = (Dl(\theta))(D\theta(\mu))(D\mu(\eta))(D(\beta)) \quad \text{[using chain rule]}
\]

(8)

Now,

\[
Dl(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \frac{1}{a(\phi)} (y - \mu)^T
\]

(9)

Again,

\[
D\theta(\mu) = (D\mu(\theta))^{-1} = \text{diag}(b''(\theta))^{-1}
\]

(10)
but, \( b''(\theta_i) = \frac{Var(y_i)}{a(\phi)} \), and \( y_i \)'s are independent. Therefore,

\[
D\theta(\mu) = a(\phi)(\text{Var}(y))^{-1}
\]

(11)

\( D\mu(\eta) \) is a diagonal matrix and

\[
D\eta(\beta) = \frac{\partial (X\beta)}{\partial \beta} = X
\]

(12)

Substituting in (8) we have

\[
\nabla Dl(\beta) = X^T (D\mu(\eta))^T (\text{Var}(y))^{-1} (y - \mu)
\]

(13)

So we have

\[
I = X^T (D\mu(\eta))^T (\text{Var}(y))^{-1} (D\mu(\eta)) X
\]

(14)

We denote

\[
(D\mu(\eta))^T (\text{diag} \{V(\mu_i)\})^{-1} (D\mu(\eta)) = W
\]

(15)

The equation for Fisher’s scoring for the exponential family becomes

\[
\left( X^T \left( W^{(t)} \right) X \right) \beta^{(t+1)} = X^T W^{(t)} \left( \eta^{(t)} + \left( D\mu(\mu^{(t)}) \right) (\eta - \mu^{(t)}) \right)
\]

(16)

Denoting

\[
\eta^{(t)} + D\eta(\mu^{(t)}) (\eta - \mu^{(t)}) = z^{(t)}
\]

(17)

Where \( D\eta(\mu) = \text{diag}\{\eta^{(\mu)}\} \). So we have

\[
\beta^{(t+1)} = \left( X^T W^{(t)} X \right)^{-1} X^T W^{(t)} z^{(t)}
\]

(18)

After several times iteration until convergence achieved, we have the estimated parameter \( \beta \).

Model Output

\[
> \text{GLMs}< \text{- glm(log(income)~Year+Month+City+Dis+WDay+Size,family=Gamma,data=d0)}
\]

\[
> \text{summary(GLMs)}
\]

Call:

\[
\text{glm(formula = log(income) ~ Year + Month + City + Dis + WDay + Size, family = Gamma, data = d0)}
\]

Deviance Residuals:

\[
\text{Min 1Q Median 3Q Max}
= -0.63150 -0.11843 -0.02213 0.09111 0.65358
\]

Coefficients:

\[
\text{Estimate Std. Error t value Pr(>|t|)}
= \text{(Intercept) 0.1848800 0.0020149 91.756 < 2e-16 ***}
\]

\[
= \text{Year2008 0.0034206 0.0009461 3.616 0.000303 ***}
\]

16
Month2  0.0043017  0.0024367  1.765  0.077570  
Month3 -0.0050059  0.0023209 -2.157  0.031069 *  
Month4 -0.0015900  0.0022450 -0.708  0.478815  
Month5 -0.0065742  0.0022316 -2.946  0.003238 **  
Month6 -0.0096556  0.0023011 -4.196  2.77e-05 ***  
Month7 -0.0222338  0.0024815 -8.960 < 2e-16 ***  
Month8 -0.0108773  0.0023049 -4.719  2.44e-06 ***  
Month9 -0.0004083  0.0022916 -0.178  0.858595  
Month10  0.0017991  0.0022478  0.800  0.423546  
Month11  0.0030891  0.0022852  1.352  0.176509  
Month12  0.0025893  0.0023082  1.122  0.262021  
CityTRUE -0.0064348  0.0011733 -5.484  4.39e-08 ***  
Dis -0.0020663  0.0001220 -16.937 < 2e-16 ***  
WDayTRUE  0.0074513  0.0009930  7.504  7.54e-14 ***  
Size  0.0067513  0.0011170  6.044  1.64e-09 ***  

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

(Dispersion parameter for Gamma family taken to be 0.02869854)
Null deviance: 134.71 on 4159 degrees of freedom
Residual deviance: 115.99 on 4143 degrees of freedom
(65 observations deleted due to missingness)
AIC: 11139
Number of Fisher Scoring iterations: 4

R code

```r
rm(list=ls(all=TRUE))
## Fetch Data ##
library(RODBC)
data<-odbcConnectExcel("data 2007 and 2008.xls")
d<-sqlFetch(data,"data")
names(d)
d0<- d[!(is.na(d$income))*(1:nrow(d)),]
## Set Workdays ##
tem<-d0[,c("TStart","TStop")]
# TRUE=Holiday
# FALSE=Workday
te<-(as.POSIXlt(tem[,1])$wday==5)&(as.POSIXlt(tem[,1])$hour>=17)
| (as.POSIXlt(tem[,1])$wday==6) | (as.POSIXlt(tem[,1])$wday==0)
d0[,"WDay"]<!-te
```

17
t<-ifelse(substr(as.POSIXlt(tem[,1]),1,10) %in% setHolidays2007,TRUE,FALSE)
d0[, "HDay"]<-t
a<-numeric(0)
for (i in 1:nrow(d0)){
  if (d0$WDay[i]==F & d0$HDay[i]==F){a[i]<-FALSE}
  else{a[i]<-TRUE}
}
d0[, "holidays"]<-a
# Set Factors #
Year <- as.factor(d0$year)
Month <- as.factor(d0$month)
Size <- as.factor(d0$size)
Bloc <- as.factor(d0$Bloc)
Dis <- as.numeric(d0$dis)
# set City or Suburb #
City<-ifelse((d0$Bloc=="mn"|d0$Bloc=="rs"|d0$Bloc=="fn"|d0$Bloc=="gt"),TRUE,FALSE)
# GLMs #
GLMs<glm(log(income)~Year+Month+City+Dis+WDay+Size,family=Gamma,data=d0)
summary(GLMs)
# 1 Histogram of Income #
hist(xlab="Income",d0$income,breaks=5000,main="Histogram of Income")
# 2 Normal QQ plot of Income #
qqnorm(d0$income,main="Normal Q-Q Plot of Income")
qqline(d0$income)
# 3 Box-Cox plot #
library(MASS)
boxcox(income~Year+Month+City+dis+WDay+Size,data=d0,lambda=seq(-1,1,1/10))
# 4 normal QQ plot of log(Income) #
qqnorm(log(d0$income),main="Normal Q-Q Plot of Income")
qqline(log(d0$income))
# 5 hist log income #
hist(xlab="log(Income)",log(d0$income),breaks=5000,main="Histogram of log(income)")
# 6 Gamma QQ plot #
beta<var(log(d0$income))/mean(log(d0$income))
alpha<mean(log(d0$income))/beta
qqplot(xlab="Theoretical Quantiles",ylab="Sample Quantiles",
log(d0$income),rgamma(10000,alpha,1/beta),main="Gamma QQ Plot")
abline(0,1)
# 7 Residual Normal QQ plot #
qqnorm(resid(GLMs))
qqline(resid(GLMs))
## 8 Ellipse Graph ##
library(ellipse)
attach(d0)
dd<-cbind(Year,Month,Dis,WDay,Size,City)
dd<-na.omit(dd)
cc <- cor(dd)
cc <- order(cc[1,])
cc <- cor(dd)[cc,cc]

colors <- c("#A50F15","#DE2D26","#FB6A4A","#FCAE91","#FEE5D9","white",
            "#EFF3FF","#BDD7E7","#6BAED6","#3182BD","#08519C")
plotcorr(cc, col=colors[5*cc + 6])

## 9 Income Changes with Month ##
INCOME<-rep(0,12)
for(i in 1:12)
{  t<-d0[d0$month==i,]
  s<-mean(t$income)
  INCOME[i]<-s
}
plot(INCOME,xlab="Month",type="b")
References


