STATISTICS
Master's Thesis
Dalarna University 2008

How Do Car-Sizes Influence the Booking?

---A research on the Stockholm Bilpool in year 2007

Author: **GuoSheng Yang**

Registration No.: **781001-T051**

Supervisor: **Mikael Möller**

Abstract

In many countries the increase of the number of automobiles has become a burden for city environment and road transport system, especially in the big cities. In order to solve the problem, car sharing may be a good way. Stockholm is the biggest city of Sweden. We analyze the data, for year 2007, of Stockholm bilpool (A car sharing organization). We want to know if the car-size is an important factor for booking frequency. Using Generalized Linear Model (GLIM) and generalized linear mixed model (GLMM), we got the result that car-size have negative effect to member's booking frequency. Car-size has positive influence to long distance booking and cold day has negative effect.

KEY WORDS: Car-sharing; Generalized linear mixed model; R

Contents

1. Introduction

1.1 What is car sharing?

Car sharing is an organization where a number of persons share the use of cars.

The following is an official definition of car sharing given in the document of Swedish National Road Administration: “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.”

Car sharing works differently from car rental. First of all, you should be a member of car sharing association just pay the entrance fee and annual fee. For every booking, you should pay the booking fee and pay the time fee and distance fee for every usage. But, it appears more abstemious than car rental. It means that car sharing is usually much less cost. Maybe it dues to members need not pay for the petrol unlike car rental.

1.2 Background

In many cities of the world, road traffic is the main source of pollution. But the number of private cars on the roads, however,

continues to increase. Instead of each individual person buying her own car, a large number of people share a small number of cars that are reserved for them and used individually as required, it really be a good method which can solve those problems mentioned above. The use efficiency of each car is higher with decreased car number. So, benefits can be obtained both from cost savings and reduced impacts on the environment. By promoting car sharing, the environmental situation should become better and people’s quality of life will be improved.

At the end of the 1987, an organized form of Car sharing was founded first in central Switzerland, and shortly afterwards in Zurich, and around a year later also in Berlin.

Stockholm's 750,000 inhabitants make more than four million journeys per day. The share of public transport is quite high (55%). [1]

In actual fact, car sharing is becoming increasingly popular in Stockholm. Some of car sharing clubs are private, and some are supported by the city. The "Stockholm Car Sharing Club" with 10 cars available for members to book. Stockholm City supports 4 car sharing clubs, with 20 "clean" vehicles (electric, ethanol, biogas) available for business trips by the city staff. In addition to this, there are approximately 10 private car sharing clubs. An easy-to-use booking system is to be developed in Stockholm, allowing city administrations and municipality-run companies to make more efficient use of their

vehicles. Stockholm Bilpool is one of them and they have 11 cars available.

1.3 Why car sharing will develop?

At first, owing a private car is however a heavy economic burden to the ordinary families. Car sharing is one of the practical ways for easing this burden.

On the other hand, it is also a recommendable method for making city environment and road transport system better. People, who participate in the car sharing organization, compare with the ones who have private cars, will be more likely to choose other ways of transportation than automobile.

The automobiles belonging to Swedish car-sharing organizations are often newer and have more satisfactory environmental and safety characteristics than Swedish cars in general. [6]

Traveling with a shared car is therefore less negative to the city environment and road transport system. All in all, car sharing is an excellent solution both for society and individual.

1.4 Aim

As it is known by us, there are many factors that influence members' booking frequency. But, based on the data, and for the sake of the

bilpool’s more optimal administration of the cars, we are interesting in the car-size and period. Maybe it will help the organization to decide which type of car they should increase or decrease.

2. Data description

We used the data of Stockholm bilpool year 2007. Three types of data are offered by them.

The first type is about the members. User number, street, city and postcode are listed. There are 231 members using the cars. I deleted 11 of them without user numbers. (Showed in Appendix B1)

The second type of data is about the time fee and distance fee. They offer different fees for different regions and car types. Seven regions are sorted to three types. Seven brands of cars fall into three levels, small, medium and large. But, there is no large car in two regions and no small car in one region. (Showed in Appendix A)

The third type of data is the records for every booking which shows user number, city, car type, car plate, used time and used distance and so on. There are some records with zero distance, probably they forgot something in the car and needed to book it to be able to open the car door. These records were deleted. (Illustrate in Appendix B2).

As the annual fee and booking fee never changed in year 2007 we can not find any influence from changing these fees.

This paper focuses on how the different types of cars affect the use of cars. There are eight sorts of cars belonging to three types. The new car is small type whose brand is Toyota Yaris. Booking is counted for different brands as showed in Appendix C.

I count the number of bookings for different car-sizes and different months as showed in Figure 1 below. (Appendix E)

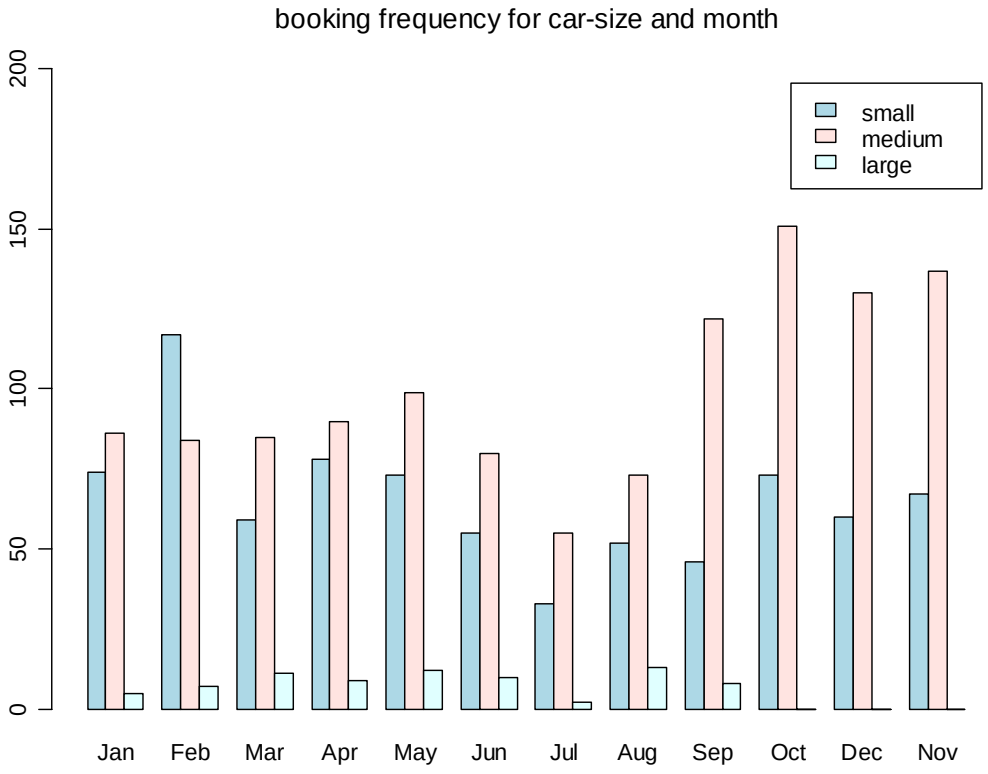

Figure 1 Booking frequency for car-size and month

There is no booking of a large car in the last three months because it was sold.

The Poisson distribution is used to model the number of events

occurring within a given time interval. The data above is the number of bookings in a month. So, a Poisson distribution should accord with the data.

The table counting booking frequency by car-size and month is a two-dimensional contingency table formed by classifying subjects by two variables. One variable determines the row categories; the other variable defines the column categories. In order to use the statistical methods usually applied to such tables, subjects must fall into one and only one row and one and only one column. [7]

3. Test

For further discussions, before we model this data with GLIM method, we should do some tests. If the row and column variables are independent, the probability of falling into a particular cell is the product of the probability of being in a particular row times the probability of being in a particular column.

Independence test:

A standard analysis of the data would be to test whether there is independence between car-size and months through a chi-square test. The corresponding GLIM approach is to model the expected number of bookings as a function of car-sizes and month. The observed numbers in

the cells are assumed to be generated from an underlying Poisson distribution. The degrees of freedom of the chi-square test equals the product of (the number of rows - 1) and (the number of columns - 1), or (r-1)*(c-1). Any disagreement between the observed and expected values will result in a large value of the chi-square statistic, because the test statistic is the sum of the squared differences. The null hypothesis of independence or homogeneity of proportions is rejected for large values of the test statistic.

In order to do the independence tests. We calculate the expectation of booking frequency. The last equality should be an approximation

$$\mu_{ij} = E(n_{ij}) = \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}}$$

Where,

$$n_{\bullet j} = \sum_{i=1}^{3} n_{ij} \qquad n_{i\bullet} = \sum_{j=1}^{12} n_{ij}$$

$$\chi^2 = \sum_{i=1}^{3} \sum_{j=1}^{12} \frac{\left(n_{ij} - \mu_{ij}\right)^2}{\mu_{ij}}$$

Hence, a difference in a cell with a larger expected cell count should be down weighted to account for this. There is another problem which the

most conservative recommendation says all expected cell counts should be 5 or more. We can see that it accords with this requirement absolutely. (Appendix D)

Using the “chisq.test()” in R, I analyzed the data and got the result as follow.

Table 1: Result of Pearson's Chi-squared test

X-squared	df	p-value
112.2092	22	4.442e-14

This test is highly significant, so we can safely conclude that the data contradict the hypothesis of independence. It means a relationship exist between car-size and month.

Test for whether booking frequencies equal for all types of cars.

$$H_0 : \qquad \mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot}$$

$$where \qquad \mu_{i\cdot} = \frac{1}{12}\sum_{j=1}^{12}\mu_{ij} \qquad i = 1,2,3$$

Table 2: Chi-squared test for given probabilities on car-size

X-squared	df	p-value
929.6449	2	< 2.2e-16

So, we get the conclusion that booking frequency is not equal for all types of cars. Different habits exist concerning car size.

Test for whether booking frequencies equal for all months.

$$H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \bullet\bullet\bullet = \mu_{\cdot 12}$$

$$\textit{where} \qquad \mu_{\cdot j} = \sum_{i=1}^{3} \mu_{ij} \qquad j = 1, 2, \cdots, 12$$

Table 3: Chi-squared test for given probabilities on month

X-squared	df	p-value
84.4825	11	1.993e-13

From the table, we can say that booking frequency of all months have significant different.

4. Modeling

In the firstly study, the members' usage did not follow the Normal distribution as shown in FIGURE 2.

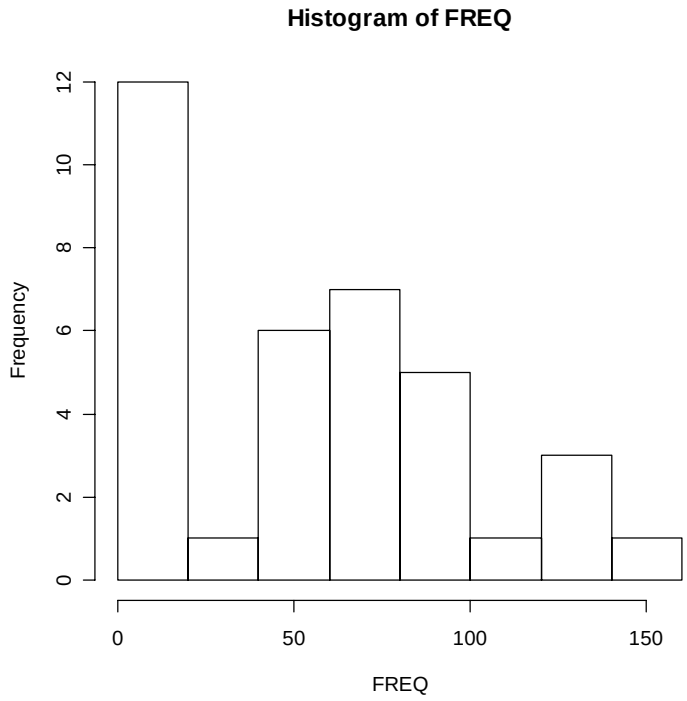

FIGURE 2: Histogram of the booking frequency

4.1 Does car-size influence the booking frequency?

Because the response is not Normal distributed and as it was mentioned before, the booking frequency in one month rather should follow a Poisson distribution than a Normal. In this situation, we should model these data with GLIM.

GLIM (Generalized Linear Model) is the generalization of GLM (General Linear Model). It can allow us to model our data using other distributions than the Normal. The choice of distribution affects the assumption we make regarding variances, since the relation between the variance and the mean is known for many distributions. For example, the

Poisson distribution has the property that the mean is equal to the variance.

I model the data using GLIM with Poisson distribution. As the canonical link of Poisson distribution is “log” link. So a first attempt to modeling these data is a generalized linear model with a Poisson distribution, a log link and a simple linear predictor. The interaction should be included because of the dependence between row and column variables.

$$\log\left(\mu_{ij}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Where α is the row variable β is the column variable and $\alpha\beta$ is the interaction. And i=1, 2, 3 j=1, 2, … , 12

Compare with two instances if car size is treated as a factor. The AIC of the model without interaction is 1223.2 which is much larger than the following model with interaction.

Table4: GLIM with Poisson distribution and “log” link.

	Estimate	**Std. Error**	**Pr(>|z|)**
(Intercept)	**4.43043**	**0.07207**	**< 2e-16 *****
MONTH	**-0.03944**	**0.01038**	**0.000146 *****
factor(CAR)2	**-0.19273**	**0.09839**	**0.050144**
factor(CAR)3	**-1.90034**	**0.22396**	**< 2e-16 *****
MONTH:factor(CAR)2	**0.09238**	**0.01340**	**5.49e-12 *****

MONTH:factor(CAR)3	**-0.07580**	**0.03612**	**0.035847 ***
Null deviance:	**1420.43 on 35 degrees of freedom**		
Residual deviance:	**154.31 on 30 degrees of freedom**		
AIC	**349.91**		

Following is the figure of QQ plot and QQ-line for the model with interaction. The residual of this model is normally distributed.

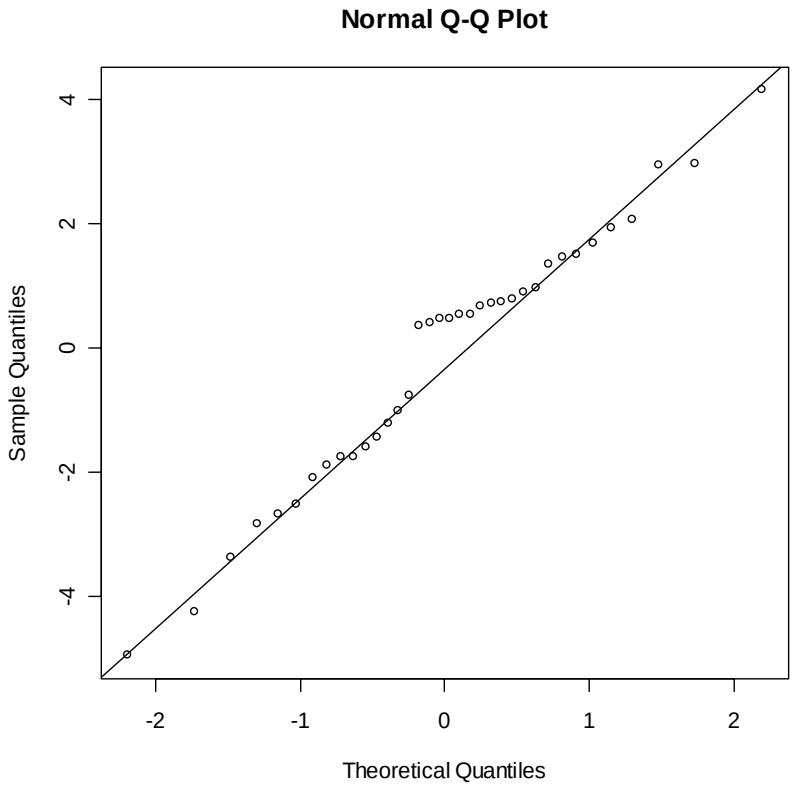

FIGURE 3: Q-Q plot of the residuals

4.2 What kind of booking did car-size affect?

Here we know that car-size has effect to the booking of cars. A question should be presented that how it influence the booking and what kind of booking it can affect. Travel distance must be known before a booking occur. So it should be important information for what type of car they will book. For analyzing this, I built a model using GLIM with normal distribution and log link as follow.

$$\text{Log(DISTANCE)=u+CARTYPE+COLD+}\varepsilon$$

Where DISTANCE is travel distance for every use, COLD is a dummy variable defined by cold period in Stockholm. Because the daylight saving time of Sweden is during April to October and it become colder in October, I put it into the cold part for dividing one year to two parts. Table follows shows the result.

Table5: estimates of the GLM

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.97655	0.07713	64.519	< 2e-16 ***
factor(type)2	0.17524	0.08945	1.959	0.050249
factor(type)3	0.49017	0.14620	3.353	0.000817 ***
factor(cold)1	-0.58520	0.08884	-6.587	5.87e-11 ***

Travel distance was affected by car-type positively. It seems that

people want to use a larger car for a longer travel.

4.3 How do car-sizes affect the booking?

For further discussion, I will use a Generalized Linear Mixed models (GLMM) to analyze how car-size influence the distance. Firstly, some knowledge should be present.

Generalized linear mixed model (GLMM) is generalized linear model with random effects **u** and conditional distributed response **y** given **u**, form as $E(y|u) = X\beta + Zu$, each element of which independently has an exponential family distribution.

$$y_i \mid u \sim indep. f_{Y_i|u}\left(y_i \mid u\right)$$

$$f_{Y_i|u}\left(y_i \mid u\right) = \exp\left\{\frac{y_i\gamma_i - b\left(\gamma_i\right)}{\tau^2} - c\left(y_i, \tau\right)\right\}$$

$$E(y_i \mid u) = \mu_i$$

$$g\left(\mu_i\right) = x_i^{'}\beta + z_i^{'}u$$

Where, random effects **u** has normal distribution, with mean equal to zero.

In R program, a package named “lme4” can be used to achieve this model. Before modeling this data, I should define the respond variable as follow.

$$y_i = \begin{cases} 1, & when\ \ record\ \ i\ \ with\ \ long\ \ distance \\ 0, & otherwise \end{cases}$$

Actually, we can define the variable as long or short distance. But I think more long distance bookings mean more income for the bilpool. So I focus on the long distance. The variable type is signed the car-size with three levels 1, 2 and 3 and a dummy variable cold as used above. User number is treated as random effects.

The model's form is as follow.

$$y_{ijk} = \alpha + \beta_i \times type + \beta_j \times cold + \mu_k + \varepsilon_{ijk}$$

Where i=1, 2, 3 and j=1, 2, … , 12

Now, the problem is how to define the long or short distance. 100 kilometers is picked as the typical one. I model the data with binomial distribution and "logit" link and the result is as follow.

Table 6: Estimates effects

Fixed effects			
	Estimate	Std. Error	Pr(>|z|)
Causal variable			
factor(type)2	0.3531	0.1429	0.0135 *
factor(type)3	0.8086	0.3559	0.0231 *
Causal variable			
(Intercept)	-0.9124	0.1735	1.45e-07 ***

factor(winter)1	-0.8180	0.1245	4.95e-11 ***

Random effects

Groups	Name	Variance	Std.Dev
usernumber	(Intercept)	1.3514	1.1625

We can see that all estimates in the model are significant. The most important thing is that both estimates of the factor type are significant. The estimate of medium car is 0.3531 with p-value 0.0135 and large car is 0.8086 with p-value 0.0231. So they both have positive effect on long distance booking. For a logistic regression, we can calculate the odds ratio by the estimate. For medium car, the odds ratio is $e^{0.3531} = 1.4235$ and the odds ratio of large car booking is $e^{0.8086} = 2.2448$. They are all bigger than 1 which means the odds of user book a car for long distance when they use a medium car or large car booking is 1.4325 or 2.2448 times than they use a small car respectively.

5. Conclusion

From the first generalized linear model, we know that car-size has an obviously negative influence to booking frequency, especially the large car does. Maybe that is why they sold the large car.

The second model shows that car-size have a positive effect to the use

distance, more bigger is longer. The effect of distance bring from cold day is negative. Because of people are tired to have a travel in cold days.

The generalized linear mix model analyzed the influence of car-size to the long distance bookings. It also presents that large car use could increase the long distance bookings.

So a bigger car in use can brought more long distance booking which can enhance more proceeds. But it would decrease booking frequency. It is really difficult to decide which type of car should be more. Maybe it is the best way to put more medium car in use.

Reference

[1]http://www.managenergy.net/products/R1379.htm

[2] Christensen Ronald, (1997), *Log-Linear Models & Logistic Regression.* Springer New York

[3] Lindsey James K, (1997), *Applying Generalized Linear Models.* Springer New York

[4] Peter Dalgaard, (2002), Introductory Statistics With R, Springer

[5] Ulf Olsson, (2002). *Generalized Linear Models: An Applied Approach*, Student litteratur AB.

[6] Per Schillander, (2003), *Make space for carsharing*, Vägverket, Publ.no.2003: 88

[7]Gerard E. Dallal (2000) *Contingency table*

Appendix

Appendix A: fees table

Bilmodell		Gubbängen, Solna, Bagartorp	Medborgarplatsen, Rosenlunds sjukhus, Gärdet	Fridhemsplan	
Avensis	timefee	**17.0**			kr/h
	distancefee	**1.95**			kr/km
Corolla, Focus, Astra	timefee	**15.5**	**22.0**	**24.0**	kr/h
	distancefee	**1.80**	**1.80**	**1.80**	kr/km
Aygo, Corsa, Fiesta	timefee			**19.0**	kr/h
	distancefee			**1.70**	kr/km

Appendix B: Some missing data

Appendix B1: Without car type or plate

Year	Month	city	location	car	plate	usernumber	hours/minutes	km
2007	4	Sydväst	Marklandsgatan			20	5:27	124

Appendix B2: Without user number or address

User number	Street	Zip code	City
128	Näckrosvägen 17	169 37	Solna
	Primusgatan 84	11267	Stockholm
58			

Appendix B3: Some deleted data with zero distance

Year	Month	city	car	plate	User number	hours/minutes	km
2007	1	Stockholm - Alla bilar	Toyota Avensis TTS674	TTS674	229	0:01	0
2007	2	Stockholm - Alla bilar	F Fiesta XNC823	XNC 823	325	0:23	0
2007	9	Stockholm - Alla bilar	F Fiesta XNC822	XNC 822	209	6:15	0

Appendix C: types of cars

small		medium		large	
Aygo	36	Focus (Ford)	558	Avensis	77
Corsa	95	Astra	114		
Fiesta	495	Corolla	520		
Toyota yaris	161				

Appendix D: Expected value of booking frequency

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
small	63.16	79.62	59.33	67.75	70.43	55.50	34.45	52.82	67.37	85.74	72.73	78.09
medium	95.66	120.59	89.86	102.62	106.68	84.07	52.18	80.01	102.04	129.87	110.16	118.27
large	6.18	7.79	5.80	6.63	6.89	5.43	3.37	5.17	6.59	8.39	7.12	7.64

Appendix E: count of different car-sizes and months

size	month											
	1	2	3	4	5	6	7	8	9	10	11	12
small	74	117	59	78	73	55	33	52	46	73	60	67
medium	86	84	85	90	99	80	55	73	122	151	130	137
large	5	7	11	9	12	10	2	13	8	0	0	0