Department of economics and society, Dalarna University
Statistics
Master's Thesis D 2008

An application of generalized linear model and mixed model

An empirical analysis of Stockholm bilpool

Author: Xin Zhao

Registration no.: 840728-T024

Supervisor: Mikael Möller

Abstract

Carsharing, as a middle choice between owning no car and holding a private car is becoming more and more popular nowadays in Europe. It is some kind of modern conception for using cars, you can use a car when you require instead of owning one. This essay is an empirical analysis of Stockholm bilpool. Using generalized linear model (GLIM) and generalized linear mixed model (GLIMM), I set up three different models to investigate how different factors can affect the booking frequence, driven distance and holding time.

Key words: Carsharing, GlIM, GLIMM, Booking frequence, Driven distance, Holding time.

Content list

1. Introduction

"The right to have access to every building in the city by private motorcar, in an age when everyone possesses such a vehicle, is actually the right to destroy the city." [1]

With the development of modern science and modern technology the automobile has become the most important mean of transportation in people's daily life. But the increasing number of vehicles has already caused many problems. Nowadays more and more people realize that to be able to drive from one place to another it is not necessary to own a car.

The idea of separating the ownership and the use of cars, carsharing, is thought to be an effective and economical solution.

Whatever you need, whenever you need it, we have a car for you.[2]

Sharing is, in some sense, almost the same as having your own car.

1.1 Definition

Carsharing means that a number of persons share the use of one or more cars. Use of a car is booked beforehand, the user pays a fee based on the distance driven and the length of time the car was made use of [3].

A carsharing enterprise is different from traditional car rentals. Firstly, they focus on a different kind of customer, the resident. The rental companies are focused on business users, tourists and the rare occasions when people move from one place to another. Car sharing is targeted towards the residents. Secondly, carsharing is much more flexible than rentals – less administrative work is needed. Thirdly, in the sharing system all users are members and they can reserve the car long before it is needed as well as on short notice.

1.2 Background

The concept of sharing a small amount of cars amongst many people was dated from 1948 in Zurich. In the early 1970s a whole system of carshare projects emerged. Switzerland and Germany were where carsharing really took place in 1980s and the first half of the 1990s.

Carsharing is a revolution in personal transportation - urban mobility for the 21st century[4]. Sharing cars with others means that you don't need to worry about the tax,

[1] Lewis Mumford, 1981
[2] http://www.citycarshare.org/carsandlocations.do
[3] the Swedish National Road Administration
[4] CarSharing.net

insurance and repairs. Fewer parking lots are needed. It can also abate the pollution of the environment and traffic jams. At the same time it will provide freedom for the car users to choose any type of car, whenever they need it. There are more than six hundred cities all around the world where people can share cars.

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 and the development of carsharing in Sweden have accelerated[5]

1.3 How it works?

The more established operations usually require an entrance fee, a monthly/yearly fee, a booking fee and a time and a distance fee. The vehicle is reserved in advance, usually over the Internet or telephone (SMS included). Most companies charge an hourly fee for the time the car is in use, plus a distance fee per driven km. Some CSO’s (Car Sharing Organisations) offer a discounted all-day rate for their cars. If a vehicle is not returned at the scheduled time, a high penalty is charged, since it may interfere with other members’ reservations. Members are responsible for leaving the vehicles on time, in the agreed parking area, clean and in good condition for the next user (available from http://en.wikipedia.org/wiki/Carsharing).

Generally speaking, carsharing can be summarized as follows:

1. You must be a member
2. Find an appropriate car.
3. Reserve it by phone or internet.
4. Use the car during the reserved time.
5. Return the car, in a proper condition, before the end of the reserved time.

1.4 Aim of the essay

Based on the data that was provided by Stockholm bilpool I will analyze what’s the real reason for the changes of booking frequence in each month, which kind of factors will affect the driven distance and the holding time for each trip.

2. Data description

2.1 The original data

The original data, from Stockholm bilpool, contains booking instances from January

[5] Make space for car-sharing! Vägverket 2003: 88E

to December, year 2007. The data also contains information on the members.

A sample of the employed booking data in February is shown in table 2.1

Table 2.1 Example of the employed data of Stockholm bilpool

month	car	user number	trip-start	trip-stop	minutes	km
2	Ford FF-WBH605	257	2007-02-24 08:10	2007-02-24 12:40	270	66
2	Ford FF-WBH605	221	2007-02-25 15:27	2007-02-25 17:25	118	30
2	Ford FF-WBH605	257	2007-02-28 15:02	2007-02-28 18:46	224	33
2	Ford FF-WBH605	221	2007-02-28 21:47	2007-02-28 22:33	46	27
2	Opel Corsa TGY278	162	2007-02-04 10:00	2007-02-04 20:00	600	183
2	Opel Corsa TGY278	91	2007-02-05 18:00	2007-02-05 21:00	180	24
2	Opel Corsa TGY278	298	2007-02-07 09:30	2007-02-07 14:00	270	107
2	Opel Corsa TGY278	234	2007-02-08 10:00	2007-02-08 18:00	480	86
2	Opel Corsa TGY278	199	2007-02-10 13:00	2007-02-11 15:00	1560	20
2	Opel Corsa TGY278	234	2007-02-13 08:00	2007-02-14 19:00	2100	167
2	Opel Corsa TGY278	254	2007-02-18 10:00	2007-02-19 07:00	1260	210

The original data contains more information than given here. The selected data is the data I will use in my analysis.

2.2 Processing data

While investigating the data a lot of inconsistent records emerged. Most of them concerned members who had booked a car but never used it (distance 0 km). All inconsistent records were deleted.

Deleted records:

Jan: 2, 3, 7, 9, 36, 38, 39, 111
Feb: 21, 80, 81, 90, 103, 109, 142, 150, 170, 172
Mar: 11
Apr: 17, 42, 87, 125, 166
May: 9, 100, 129, 132, 175
Jun: 31, 67, 89, 90, 114
Jul: 31, 55, 64, 69, 80, 96
Aug: 4, 70, 103
Sep: 13, 20, 35, 36, 48, 60, 77, 84, 91, 98, 132, 133, 152
Oct: 12, 15, 18, 99, 102, 150, 174, 189, 201, 235[6]
Nov: 5, 8, 12, 13, 62, 63, 114, 116, 143, 170, 172, 180, 203
Dec: 9, 76, 148, 153, 155, 157, 169, 178, 186, 201, 203, 205

After the data has been cleaned from inconsistencies it was designed into two

[6] Fabia (XSL 202) is not owned by Stockholms BilPool

different data sets, which can be distinguish as data set A[7] and data set B[8] (see appendices for examples).

The meanings of the variables in appendix A are as follow:

1. freq: is the number of times a car has been booked.
2. CS: is to distinguish between city or suburb[9] residents, 1=city, 0=suburb
3. holiday: is the red day in Swedish calendar for 2007.
4. night: is defined as the time between 17:00 and 06:00.

The meanings of the variables in appendix B are as follow:

1. ID: is the identification number of the user.
2. minutes: is the time per trip.
3. km: is the driven distance per trip.
4. car1, car2, car3: indicates the type of car booked[10].
5. CS: is to distinguish between city or suburb residents, 1=city, 0=suburb
6. holiday: is the red day in Swedish calendar for 2007.
7. night: is defined as the time between 17:00 and 06:00.

3. Methodology

According to the aim of this essay I will analyse data set A to show how the booking frequence is affected by other factors, and also analyse data set B to find out which kind of factors will affect the driven distance and holding time.

3.1 Generalized linear model

Generalized linear model (GLIM) is a flexible generalization of ordinary least squares regression. In general linear models (GLM), there are several assumptions such as normality, homoskedasticity and linearity. However, in GLIM we will give up such kind of assumptions: drop the normality in favor of the exponential family of distributions; abandon the homoskedasticity in favor of a known function, which is called variance function, and explain how the individual variation depend on the respective mean; throw away the linearity assumption in favor of a known function which is called the link function and then translate the nonlinearity into a function of linear relationships which we call linear predictors. The structure of the GLIM model may be displayed as in following graph:

[7] The details of data set A is shown in appendix A.
[8] The details of data set B is shown in appendix B.
[9] The postal code of Stockholm is absolute different from other areas, so I just simply thought Stockholm is city side all others are suburb. Stockholm has zipcode 1005 to 12033, exclude 10504.
[10] There are three different kinds of types. 1=large=Avensis, 2=medium=Focus, Astra, Corolla, 3=small=Corsa Fiesta Yaris. Aygo was bought in 2007 and replaced Yaris. Corsa has been replaced by Ford Fiesta. Avensis was replaced by Corolla. (Hence no big cars anymore. But include big cars in the analysis)

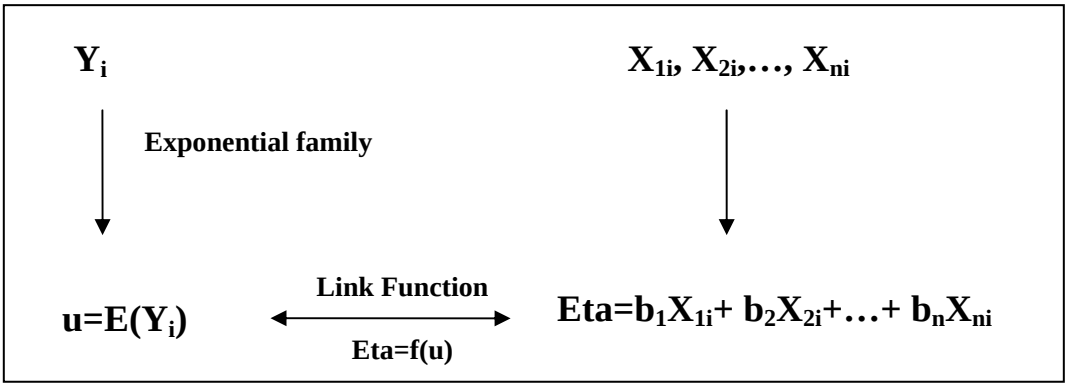

Figure 3.1

From the flow chart above, we see that the response variable belongs to the family of exponential distributions such as Gaussian distribution, Poisson distribution, Gamma distribution and so on. Hence the distribution for the response variable may be adapted to the data. For example, if the data is continues then we can select the Gaussian or Exponential distribution for the response variable Y, or if the data is a count, then we can choose the Poisson or Binomial distribution. Then we can compute the expected value u of the response variable Y. On the other side, we can choose several independent variables which are potential factors to the response variable. After computing the linear predictor Eta, we employ the corresponding link function to get the nonlinear relationship between the expected value and the linear predictor.

In this essay, we apply the GLIM model to explore several factors which potentially affect the booking frequence given data set A. More details will be discussed in next section.

3.2 Repeated measures data

Repeated measures data is a kind of data set which contains multiple observations per case and it is usually obtained from multiple measurements of a response variable. Such multiple measurements can be obtained from each experimental unit over time or under multiple conditions. In this thesis, data set B is obtained from multiple measurements over time. Thus, it belongs to repeated measures data.

There are several methods that can be employed to analyze repeated measures data, such as separate analyses at each time point, fit multilevel growth curve, univariate and multivariate analyses of time contrast variables, and mixed model methodology. In this paper, mixed-effect model will be employed to deal with the data analysis.

3.2 Generalized Linear Mixed Model

In linear model or generalized linear model, we assume that all observations are

independent. However in empirical problems, observations are always measured over repeated occasions, just as the repeated measures which we have mentioned above. Some times it is the case that observations are clustered. In this situation, the estimation of regression parameters become to biased and inconsistent. Mixed model may solve such kind of problems.

In my essay, the monthly data set is concerned about the driven distance and the time used for different members during the year. Thus, there must be some relationship among each customer's behavior in the whole year, and it is obvious that the mixed model is necessary for this data set in the essay.

Next, I'd like to introduce some basics about the Linear Mixed model (LMM) and Generalized Linear Mixed model (GLIMM) which will be used in this paper. Mixed-effects models provide a flexible and powerful tool for the analysis of grouped data, which arise in many areas as diverse as agriculture, biology, economics, manufacturing, and geophysics, (see Jose C.P. and Douglas M. B. (2000)). Mixed models are models where some of the independent variables are assumed to be fixed, i.e. chosen beforehand, while others are seen as random sampled from some population or distribution (see Ulf O. (2002)). Thus mixed models are constructed by two parts: a fixed part and a random part. A common linear mixed model can be written as:

$$Y = X\beta + ZU + \varepsilon$$

Here Y is a (N$\times$1) vector of observation; X is a (N$\times$P) design matrix for fixed effects; β is a (P$\times$1) vector of fixed, unknown parameter; U is a (Q$\times$1) vector of random effects; Z is a (N$\times$Q) design matrix for random effects; ε is an (N$\times$1) vector of residual random errors.

4. Models

4.1 GLIM for Booking Frequence

4.1.1 Model description:

I apply data set A to analyze how the booking frequence depends on month, holiday, night and CS. Because the response variable Booking frequence can be seen as count, I assume that it has an Poisson distribution, and I will test this in the next section of model diagnose. I set Log link as the link function and use month, holiday, night and CS as independent variables. The linear predictor can be expressed as follows:

$$\eta = \beta_1 month + \beta_2 CS + \beta_3 holiday + \beta_4 night + \beta_5 night \cdot holiday + \beta_6 CS \cdot holiday + \beta_7 night \cdot CS + \beta_8 night \cdot holiday \cdot CS$$

4.1.2 Fitting result and Model diagnose

The fitting result is showed in the table 4.1.

Table 4.1

```
glm(formula = freq ~ factor(month) + cs + holiday + night + cs *
    night + night * holiday + night * cs * holiday - 1, family = poisson,
data = d)

Deviance Residuals:
   Min      1Q    Median    3Q      Max
-3.0160  -0.7957  -0.2102   0.5484   2.5637

Coefficients:
                 Estimate   Std. Error  z value    Pr(>|z|)
factor(month)1    3.62472    0.08978    40.376    < 2e-16 ***
factor(month)2    3.58722    0.09107    39.389    < 2e-16 ***
factor(month)3    3.52819    0.09317    37.867    < 2e-16 ***
factor(month)4    3.60615    0.09041    39.885    < 2e-16 ***
factor(month)5    3.67847    0.08797    41.815    < 2e-16 ***
factor(month)6    3.42132    0.09718    35.207    < 2e-16 ***
factor(month)7    2.87478    0.12211    23.542    < 2e-16 ***
factor(month)8    3.32676    0.10094    32.958    < 2e-16 ***
factor(month)9    3.48680    0.09469    36.822    < 2e-16 ***
factor(month)10   3.71276    0.08685    42.750    < 2e-16 ***
factor(month)11   3.50771    0.09392    37.347    < 2e-16 ***
factor(month)12   3.48680    0.09469    36.822    < 2e-16 ***
cs                0.37844   0.06490     5.831    5.51e-09***
holiday          -0.73919   0.08795    -8.404     < 2e-16 ***
night            -1.25527    0.10617   -11.823    < 2e-16 ***
cs:night         -0.47024    0.15033    -3.128     0.001760 **
holiday:night    -0.86151    0.24492    -3.518     0.000436***
cs:holiday        0.07972    0.11296    31.909     0.00354 ***
cs:holiday:night -0.12767   0.35302    -0.362     0.717611


(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 8246.96  on 96  degrees of freedom
Residual deviance:  109.57  on 77  degrees of freedom
AIC: 540.18
```

From table 4.1, we see that all coefficients are significant except the interactive term of cs, holiday and night. And the AIC is 540.18.

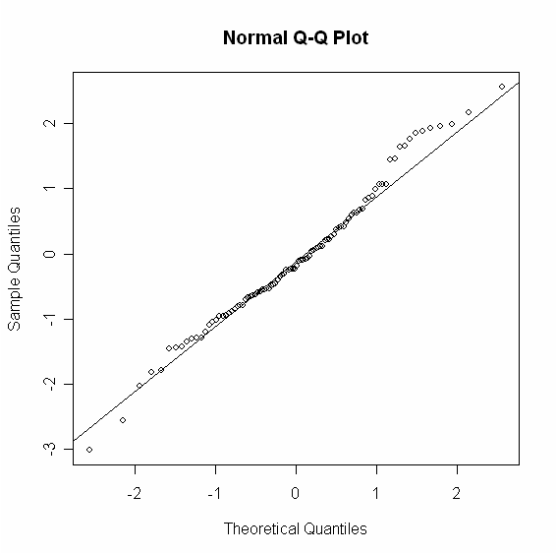

Figure 4.1

From the qqplot we see that the residuals are close to the normal distribution. In other words, the Poisson distribution does fit our data quite well.

4.1.3 Summary of results from the model

From table 4.1 we conclude that there are significant differences for the booking frequence for the 12 months. All p-values are extremely small.

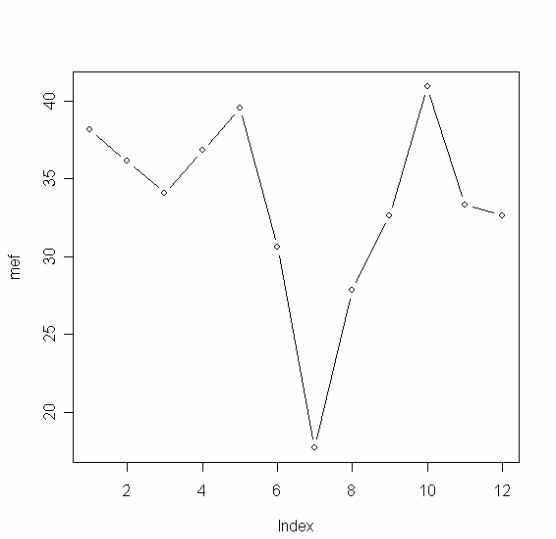

Figure 4.2

Given figure 4.2, we conclude that the averages of booking frequence are on different levels. The booking frequence in March, June, August, September, November and December stay in the middle level; in January, February, April, May and October stay in the high level; it is obvious that July stay in the lowest level.

Furthermore, from table 4.1 we can see that CS has positive effect to the booking

frequence. It is to say, the booking frequence for people who live in city is greater than those who live in the suburbs. Oppositely, holiday and night have both a negative effect. We can say that the booking frequence will decrease on holidays and nights.

4.2 GLIMM for driven distance

4.2.1 Model description:

In this model, I want to analyze which factors affect the distance driven with the aid of data set B, thus I choose driven distance as response variable and check its distribution.

First I plot the histogram and the normal qqplot and they are seen in figure 4.3. It is seen that a normal distribution is not good.

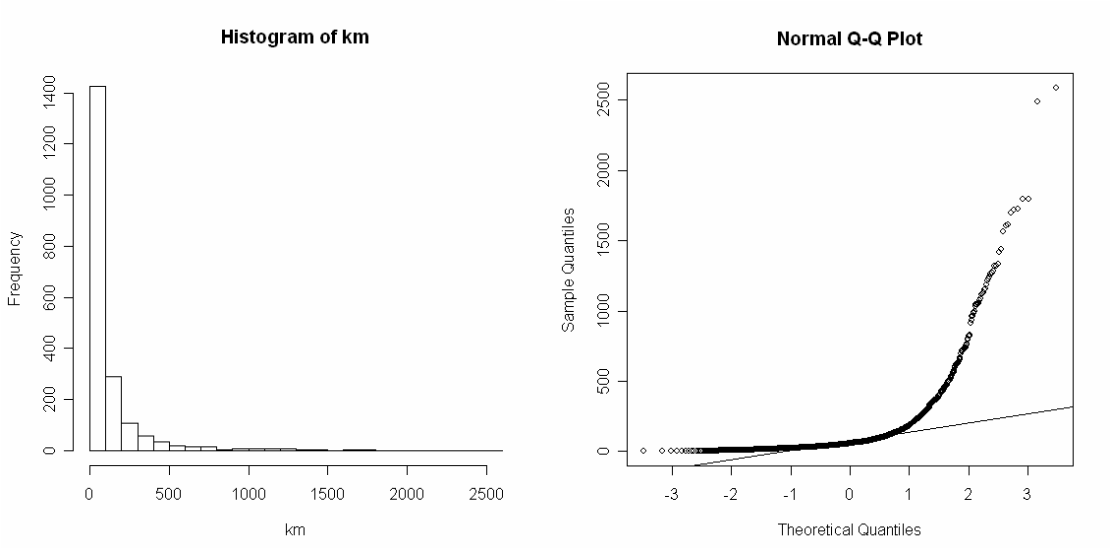

Figure 4.3

Thus, I try to use “log” transform to normalize driven distance, and the results of normalization is shown in figure 4.4.

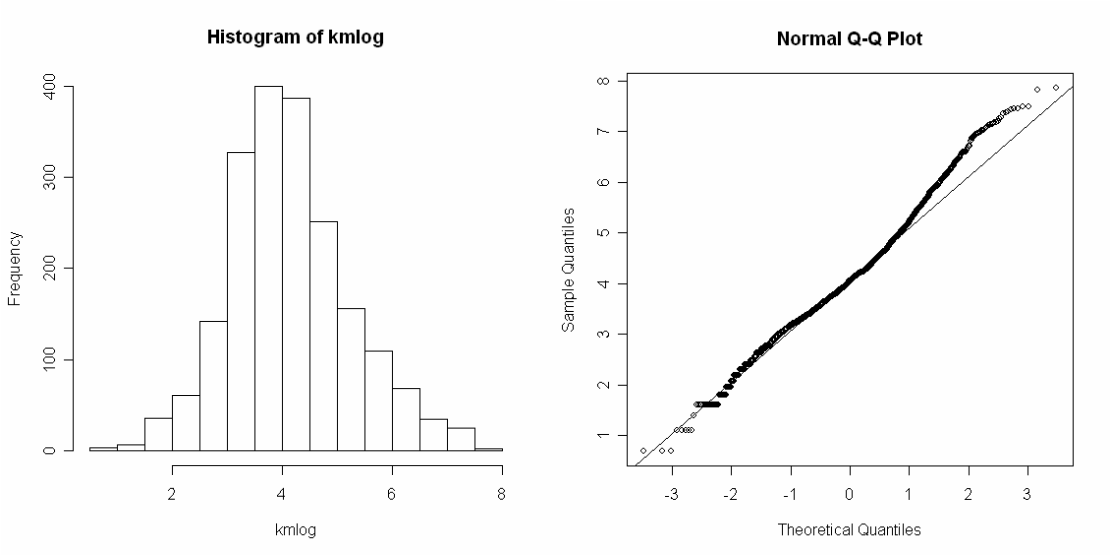

Figure 4.4

After the normal transition, I set kmlog as response and month, carsize, holiday, night and CS as the main factors which affect the driven distance, ID is a random effect. The linear predictor is:

$$\eta = \beta_1 month + \beta_2 CS + \beta_3 holiday + \beta_4 night + \beta_5 car1 + \beta_6 car2 + \beta_7 car3$$
$$+ \beta_8 night \cdot holiday + \beta_9 CS \cdot holiday + \beta_{10} night \cdot CS + \beta_{11} night \cdot holiday \cdot CS$$

4.2.2 Fitting result and Model diagnose

The fitting results are showed in table 4.2.

Table 4.2

```
Linear mixed-effects model fit by maximum likelihood
Data: d
AIC  BIC logLik
 NA   NA    NA
Random effects:
 Formula: ~1 | ID
         (Intercept)  Residual
StdDev:   0.4934518  0.9437274
```

```
Fixed effects: kmlog ~ factor(Month) + car1 + car2 + car3 + cs
+ holiday + night +     cs * night * holiday - 1
                  Value   Std.Error   DF   t-value   p-value
factor(Month)1   4.480654 0.21455521 1852 20.883452  0.0000
factor(Month)2   4.367724 0.21450065 1852 20.362290  0.0000
factor(Month)3   4.693855 0.21252088 1852 22.086558  0.0000
factor(Month)4   4.611757 0.21251999 1852 21.700345  0.0000
factor(Month)5   4.727864 0.21061464 1852 22.447936  0.0000
factor(Month)6   4.863306 0.21592856 1852 22.522752  0.0000
factor(Month)7   5.474645 0.22443646 1852 24.392850  0.0000
factor(Month)8   5.005291 0.21248967 1852 23.555455  0.0000
factor(Month)9   4.652369 0.21514723 1852 21.624119  0.0000
factor(Month)10  4.541344 0.21607538 1852 21.017405  0.0000
factor(Month)11  4.587291 0.21729467 1852 21.110921  0.0000
factor(Month)12  4.489534 0.21629035 1852 20.756980  0.0000
car1            -0.259134 0.13731245 1852 -1.887184  0.0593
car2            -0.503429 0.17419702 1852 -2.890000  0.0039
car3            -0.575597 0.17804415 1852 -3.232889  0.0012
cs               0.077678 0.11551018  137  0.672479  0.5024
holiday          0.175159 0.08787220 1852  1.993342  0.0464
night           -0.267115 0.09640669 1852 -2.770715  0.0056
cs:night        -0.042068 0.12871047 1852 -0.326844  0.7438
cs:holiday      -0.012787 0.11186038 1852 -0.114313  0.9090
holiday:night    0.207551 0.22057203 1852  0.940966  0.3468
cs:holiday:night -0.092681 0.29414828 1852 -0.315082  0.7527
```

From table 4.2, we can see that the coefficients are significant except CS and all the interactions.

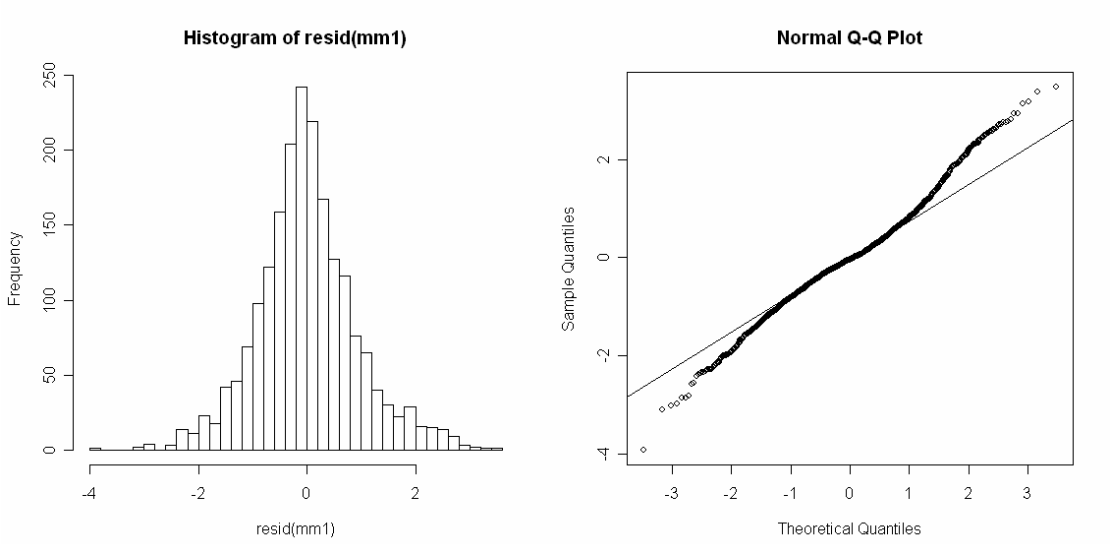

Figure 4.5

From figure 4.5, we see that the histogram resembles a normal density and the qqplot

indicates some deviation from the normal distribution. All in all we accept the assumption of a normal distribution even if it is not perfect.

4.2.3 Summary of results from the model

A plot for each month reveals the change in distance driven. From figure 4.6, we see that the distance driven in July is highest of all months. January, February and December are all at a lower level. Thus, we conclude that distance driven tends to be longer than in winter.

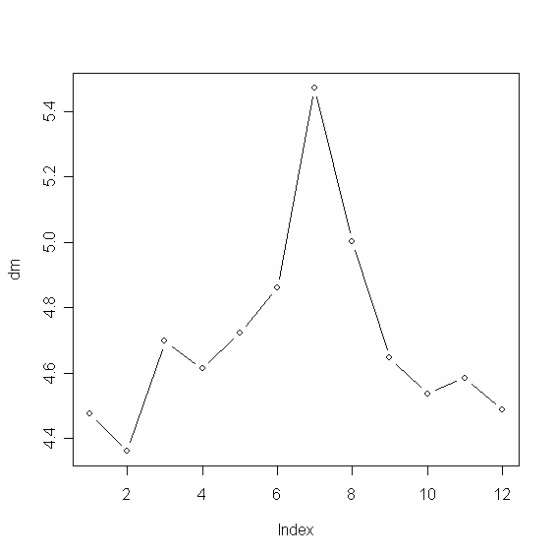

Figure 4.6

Secondly, from the other coefficients we see that the holiday factor is positive but the night factor is negative. Thus we come to the conclusion that holiday trips tends to be long and night trips short.

Thirdly, concerning car size, from table 4.2 we see that the coefficients for the different types of car are different. The larger the car size the greater the coefficient is. Hence those people who book a large car tend to take a longer trip.

At last, we see that the coefficient of CS is not significant, thus we infer that there is no difference between city and suburb residents when it comes to distance driven.

4.3 GLIMM for holding time

4.3.1 Model description:

Next I analyze, with the aid of data set B, which factors affect the holding time. thus I choose holding time as response variable and check its distribution.

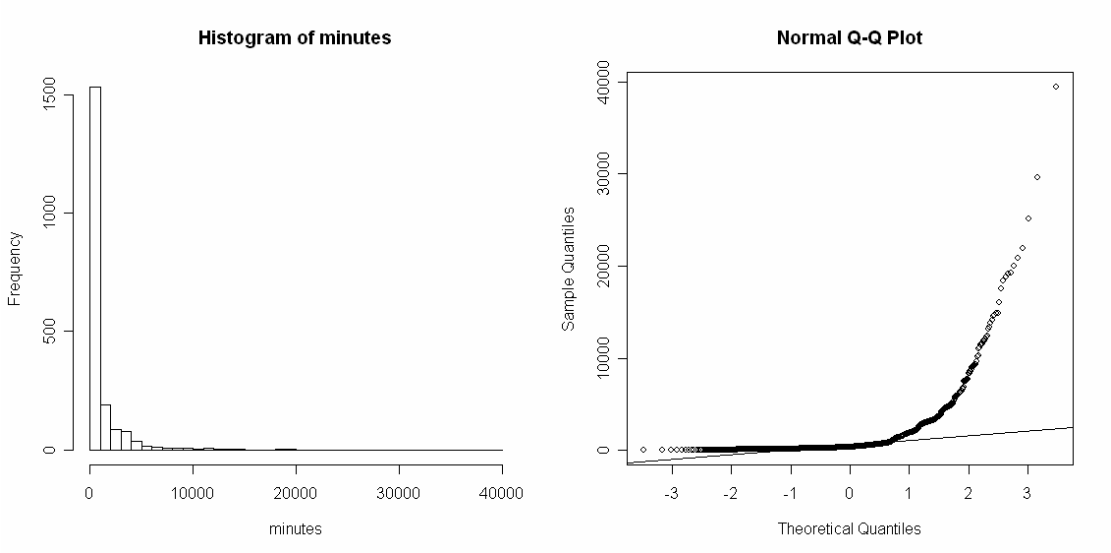

Figure 4.7

These graphs give the same information as the graphs for distance driven. Thus I decide to use the same transformation method to solve it, however, the fitting result is not very good. Then I try to apply Box-Cox transformation to correct the models

The transformation function is

$$\min = \frac{Minutes^{\lambda} - 1}{\lambda}, \ \ where\ \lambda = -0.19$$

And the transformation result is showed in figure 4.8:

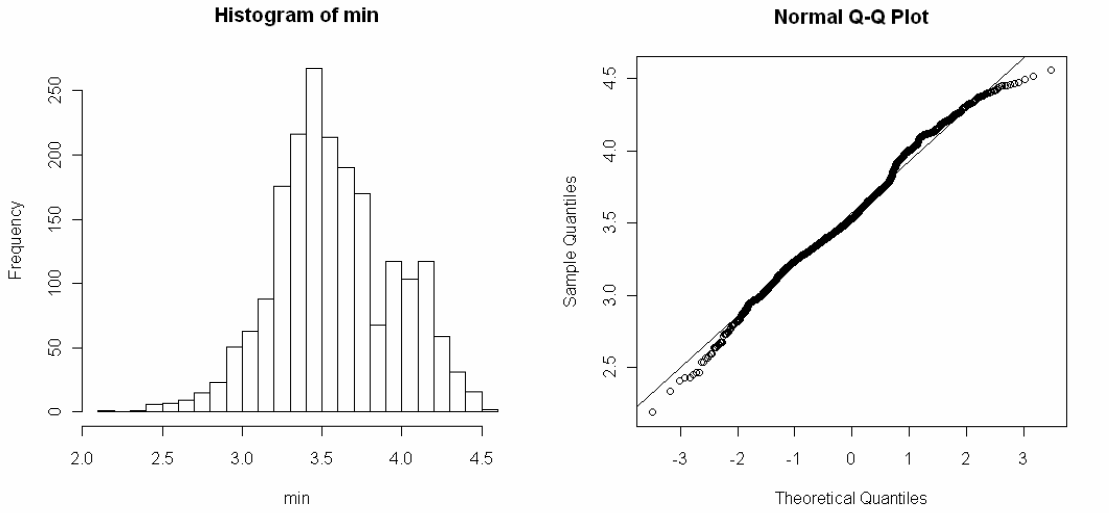

Figure 4.8

We set min as response variable and select the Gaussian distribution and canonical link as link function. In the same way as before I set month, carsize, holiday, night and CS as the main factors and ID as a random effect. The linear predictor is:

$$\eta = \beta_1 month + \beta_2 CS + \beta_3 holiday + \beta_4 night + \beta_5 car1 + \beta_6 car2 + \beta_7 car3 + \beta_8 night \cdot holiday + \beta_9 CS \cdot holiday + \beta_{10} night \cdot CS + \beta_{11} night \cdot holiday \cdot CS$$

4.3.2 Fitting result and Model diagnose

The fitting results are showed in table 4.3

Table 4.3

```
Linear mixed-effects model fit by maximum likelihood
Data: d
AIC BIC logLik
 NA  NA    NA
Random effects:
Formula: ~1 | ID
        (Intercept)  Residual
StdDev:   0.1874630   0.3071502
Fixed effects: min ~ factor(Month) - 1 + car1 + car2 + car3 + holiday+
night +     cs + cs * holiday * night
                  Value      Std.Error    DF    t-value    p-value
factor(Month)1    3.869236    0.07348828 1852   52.65105    0.0000
factor(Month)2    3.772852    0.07344984 1852   51.36637    0.0000
factor(Month)3    3.923770    0.07281905 1852   53.88383    0.0000
factor(Month)4    3.880966    0.07277250 1852   53.33011    0.0000
factor(Month)5    3.901538    0.07224434 1852   54.00476    0.0000
factor(Month)6    3.909452    0.07389835 1852   52.90310    0.0000
factor(Month)7    4.110941    0.07652368 1852   53.72115    0.0000
factor(Month)8    3.973638    0.07282199 1852   54.56646    0.0000
factor(Month)9    3.852767    0.07366487 1852   52.30128    0.0000
factor(Month)10   3.833344    0.07401352 1852   51.79248    0.0000
factor(Month)11   3.853145    0.07431119 1852   51.85148    0.0000
factor(Month)12   3.844860    0.07398501 1852   51.96809    0.0000
car1             -0.094011    0.04800962 1852   -1.95817    0.0504
car2             -0.243667    0.05900311 1852   -4.12972    0.0000
car3             -0.267095    0.06025241 1852   -4.43293    0.0000
holiday          -0.008272     0.02875437 1852   -0.28768   0.7736
night            -0.132490    0.03155891 1852   -4.19817    0.0000
cs               -0.017718    0.04136590  137   -0.42833    0.6691
holiday:cs        0.023293    0.03659323 1852    0.63655    0.5245
night:cs         -0.046794    0.04209299 1852   -1.11167    0.2664
holiday:night     0.174981   0.07199522 1852    2.43046    0.0152
holiday:night:cs -0.132438   0.09595298 1852   -1.38024    0.1677
```

From table 4.3, we see that the coefficients are significant except CS, holiday and the interaction terms.

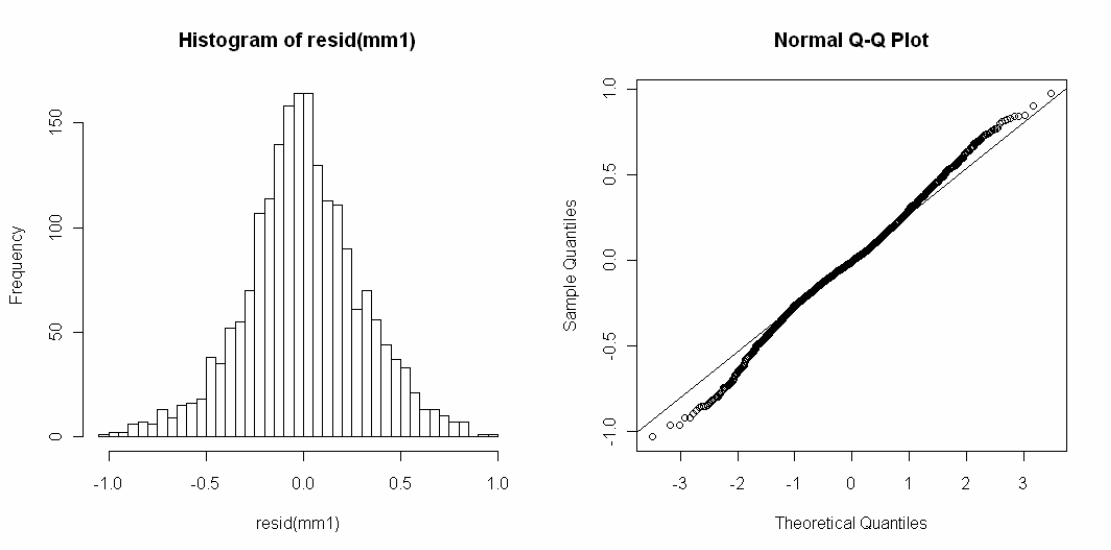

Figure 4.9

From figure 4.9 we see that the qqplot of residuals is close to normal distribution, thus we can say that this model is acceptable.

4.3.3 Summary of results from the model

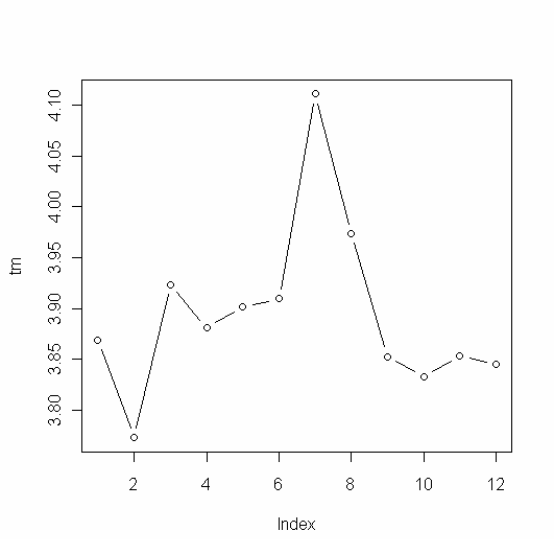

Figure 4.10

We plot the coefficients for each month to observe the changes of different levels for holding time. From figure 4.10 we get almost the same result as that for the level of distance. In July the holding time is higher than in any other month. January, September, October, November and December are all at a lower level. Thus we conclude that, the holding time for people who book a car, tends to be longer in

summer and it is the opposite in winter. Furthermore, we see that February has the lowest level among all months. However, I can't give any explanation, to do that more knowledge of how the booking system works is needed.

Secondly, concerning car size, the results are similar to previous result. From table 4.3 we can see that the coefficients of different types of car are different, and the larger the car the greater the coefficient will be. People who book larger size car tends to have a longer time.

Thirdly, we see that the effect of night is negative, thus we can infer that people who book a car for the night will hold the car for a shorter period.

At last, the significance of coefficient of CS and holiday don't pass the test, which is there are no difference for holding time for city/suburb and holiday/workday.

5. Conclusion

In this essay, I apply generalized linear model and generalized linear mixed model to analyze two kinds of data set. I investigate the factors that affect the booking frequence on different levels. Then, study how these factors affect the driven distance and the holding time.

First, after fitting the GLIM model, I found that the booking frequencies are different in different month, where booking frequence in July is at the lowest level. Furthermore, CS has a positive effect to the booking frequence. Thus we can conclude that the booking frequence in July is lower than other month; people who live in city are more likely to book a car.

Secondly, after fitting the GLIMM model, we found that driven distance and holding time both are different for each month and are both at the highest level in July. At the same time, car size has a common effect on driven distance and holding time. It is to say that the driven distance and holding time for those persons who booked a bigger car will lead to longer distances and times. Furthermore, holiday has positive effect to driven distance; night a negative effect to both holding time and distance driven. Thus we can conclude that people who booked a car in the summer and the holiday will hold it for a longer period and drive it for a longer distance; if people book a car at night, then they will hold the car for a short period and drove it for a shorter distance.

6. Subjective suggestions

According to the conclusions above, I'd like to give some subjective suggestions to Stockholm bilpool.

In July, people tends to book a car for spending there vacation, which is indicated by the low booking frequence, the longer holding time and the driven distance. Thus, the bilpool can appropriate reduce the holding fee and the distance fee[11] during July to increase the booking frequence in order to get more income.

For the night and the holiday users, give them a discount to enhance the booking frequence during such period.

7. Reference

Jose C.P. and Douglas M. B., 2000. *Mixed-Effects Models in S and S-PLUS*, Springer-Verlag New York

Ulf O., 2002. *Generalized Linear Models - an applied approach*, Lund, Sweden

Robert A. M., William L. S., Walter W. S., 1991. A Unified Approach to Mixed Linear Models. The American Statistician, Vol. 45, No. 1, 54-64.

[11] The member have to pay the price per hour as well as the price per km

Appendix A:

An example of data set A:

freq	month	cs	holiday	night
1	1	1	1	1
1	2	1	1	1
2	3	1	1	1
2	4	1	1	1
3	5	1	1	1
4	6	1	1	1
1	7	1	1	1
2	8	1	1	1
1	9	1	1	1
2	10	1	1	1
0	11	1	1	1
1	12	1	1	1
33	1	1	1	0

Appendix B:

An example of data set B

ID	Month	minutes	km	car1	car2	car3	cs	holiday	night
9	2	327	39	0	1	0	1	1	0
9	2	203	44	0	0	1	1	0	0
9	2	383	27	0	0	1	1	1	0
9	3	314	37	0	1	0	1	1	0
9	5	323	55	0	1	0	1	0	0
9	5	1884	214	0	1	0	1	1	0
9	9	1778	94	0	0	1	1	1	0
9	9	515	216	0	1	0	1	0	0
9	9	95	20	0	1	0	1	0	1
9	9	274	35	0	1	0	1	0	0
9	10	476	34	0	1	0	1	0	0
9	10	2033	228	0	1	0	1	0	0
9	12	725	202	0	1	0	1	0	0
9	12	3533	799	0	1	0	1	0	0
9	12	605	14	0	1	0	1	0	0
9	12	704	188	0	1	0	1	0	0
9	12	366	41	0	0	1	1	0	0
11	3	1886	320	0	1	0	0	1	0
11	4	701	159	0	1	0	0	0	0
11	5	472	36	0	1	0	0	1	0
11	5	4991	355	0	1	0	0	1	0
11	6	459	64	0	1	0	0	1	0