MA 684 Midterm Project

Mengyun Li

November 29, 2017

1. Introduction

I use the allstate insurance purchase history as the dataset of this project. As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. This is represented in this data as a series of rows that include a customer ID, information about the customer, information about the quoted policy, and the cost. The task of this project is to predict the purchased coverage options using a limited subset of the total interaction history.

2. Description of the data

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   time = col_time(format = ""),
##   state = col_character(),
##   car_value = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

customer_ID	shopping_pt	record_type	day	time	state	location	group_size	homeowner
1e+07	1	0	0	08:35:00	IN	10001	2	0
1e+07	2	0	0	08:38:00	IN	10001	2	0
1e+07	3	0	0	08:38:00	IN	10001	2	0
1e+07	4	0	0	08:39:00	IN	10001	2	0
1e+07	5	0	0	11:55:00	IN	10001	2	0
1e+07	6	0	0	11:57:00	IN	10001	2	0
1e+07	7	0	0	11:58:00	IN	10001	2	0
1e+07	8	0	0	12:03:00	IN	10001	2	0
1e+07	9	1	0	12:07:00	IN	10001	2	0
##Variable Des	cription							
There are tota	l 25 variables	for this data	and w	ith the var	iables d	escriptions	as follow:	

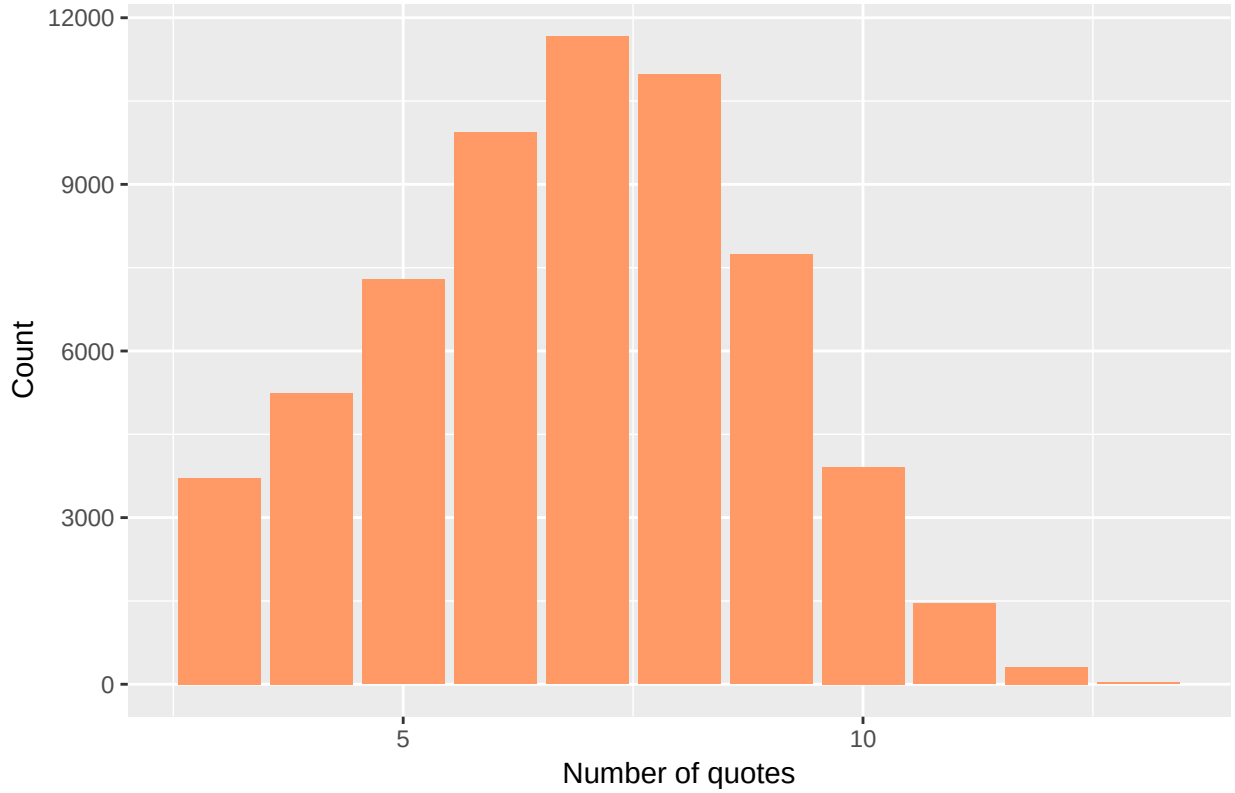
customer_ID - A unique identifier for the customer shopping_pt - Unique identifier for the shopping point of a given customer record_type - 0=shopping point, 1=purchase point day - Day of the week (0-6, 0=Monday) time - Time of day (HH:MM) state - State where shopping point occurred location - Location ID where shopping point occurred group_size - How many people will be covered under the policy (1, 2, 3 or 4) homeowner - Whether the customer owns a home or not (0=no, 1=yes) car_age - Age of the customer's car car_value - How valuable was the customer's car when new risk_factor - An ordinal assessment of how risky the customer is (1, 2, 3, 4) age_oldest - Age of the oldest person in customer's group age_youngest - Age of the youngest person in customer's group married_couple - Does the customer group contain a married couple (0=no, 1=yes) C_previous - What the customer formerly had or currently has for product option C (0=nothing, 1, 2, 3,4) duration_previous - how long (in years) the customer was covered by their previous issuer A,B,C,D,E,F,G - the coverage options cost - cost of the quoted coverage options

Explanation with the Customer ID

As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. For example, from the data above, a customer with ID 10000000 recieved nine quote and purchased the last one.

customer_ID	shopping_pt	record_type	day	time	state	location	group_size	homeowner	car_age	car
10000000	9	1	0	12:07:00	IN	10001	2	0	2	g
10000005	6	1	3	09:09:00	NY	10006	1	0	10	e
10000013	4	1	4	09:31:00	WV	10014	2	1	3	d

Figure 2.1: Number of quotes until purchase



Each customer has many shopping points, where a shopping point is defined by a customer with certain characteristics viewing a product and its associated cost at a particular time. The data related to customers have these characteristics: 1) Some customer characteristics may change over time (e.g. as the customer changes or provides new information), and the cost depends on both the product and the customer characteristics. 2) A customer may represent a collection of people, as policies can cover more than one person. 3) A customer may purchase a product that was not viewed.

3. EDA

Before I fit the model, I did some data visualization to help me understand the data.

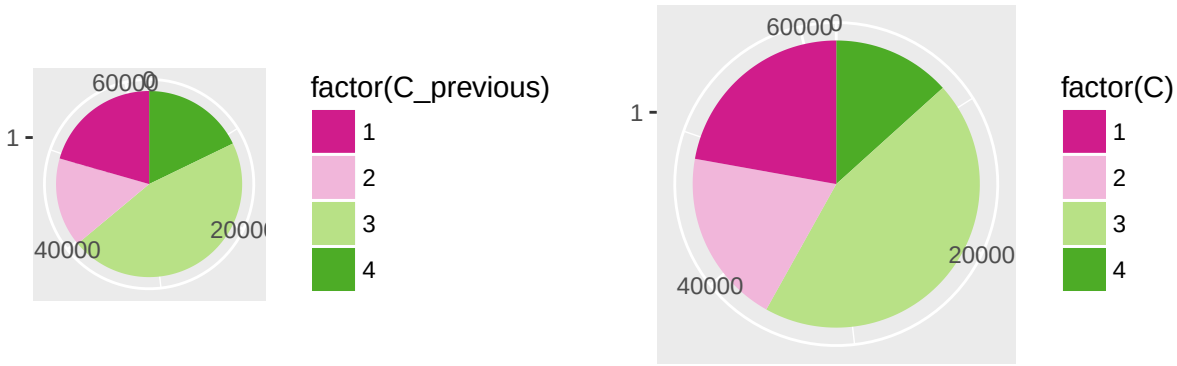
3.1 Purchased/Unpurchased option choose.

Figure 3.1: Percentage of purchase for each option

From the figure above I found that customers has not much difference in choosing each options. (No preferece in certain one option.) Also most of customers need more than five quotes then finally purchase.

3.2 The comparasion between previous option C and option C

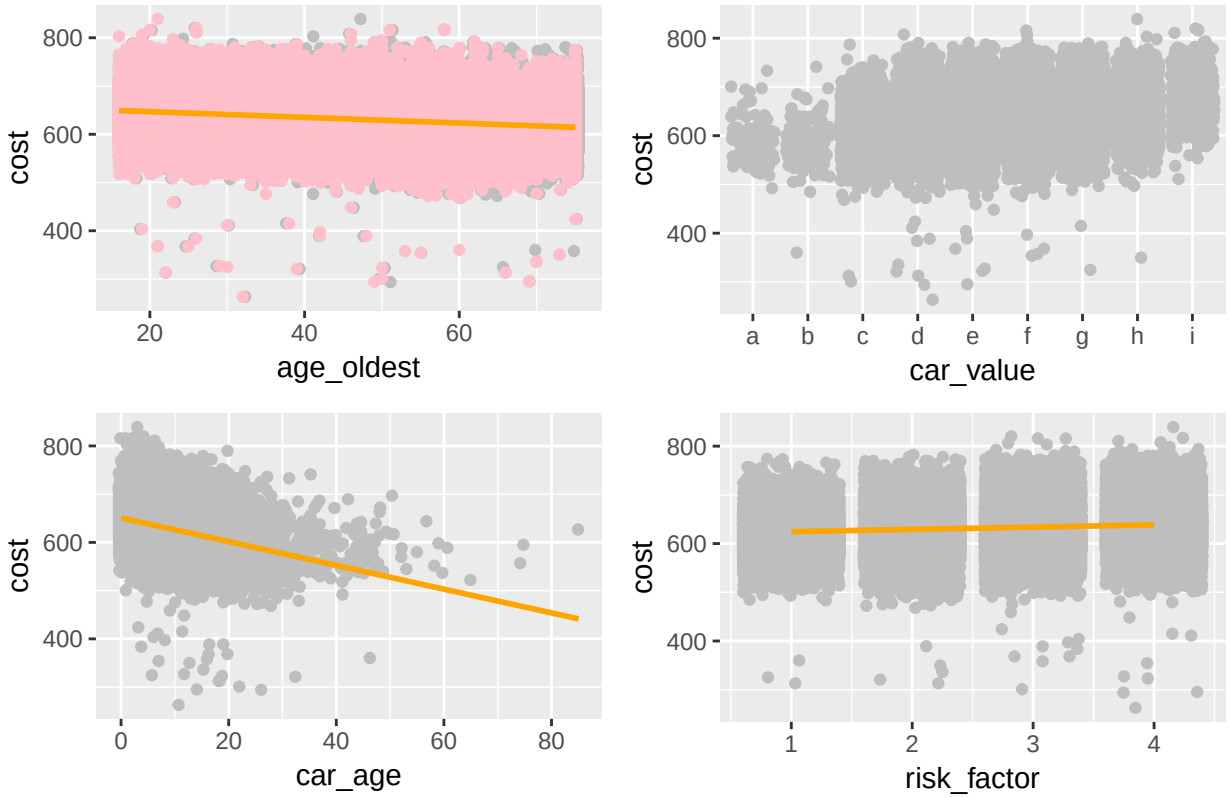
Figure 3.2: Percentage of each option in option C



From Figure 3.3 I found that there are not much difference in the amount of each option in category C, the number of people who choose option 3 decreased while the number of person who choose option 2 increased.

3.3 The relation ship between the oldest age in one group and the cost.

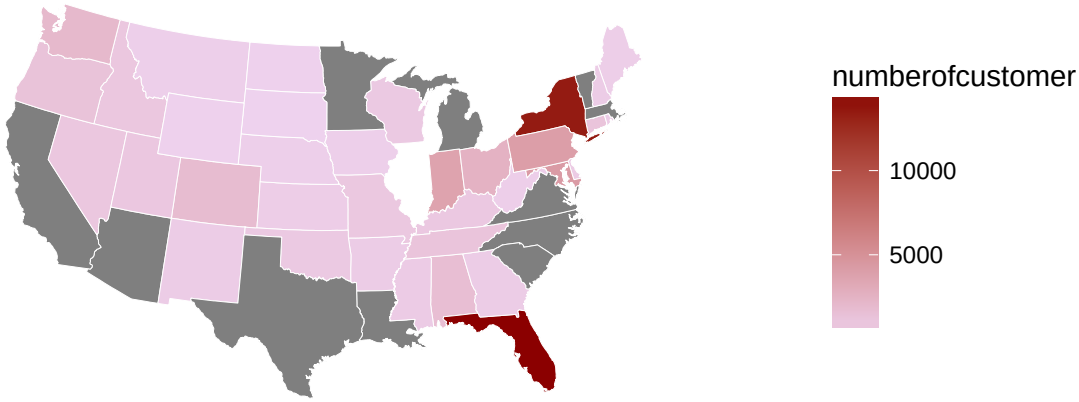
Figure 3.4: Percentage of each option in option C



From Figure 3.4 I found that the insurance cost will have lighly increase while the group age, car_value and risk factor increased. Also the cost has very obvious decreasing while the car_age is increased.

3.4 Number of customers in each state

Figure 3.4 Chloropleth map of amount of customer in each states



From figure 3.4 we can see there is a large difference between the number of customers in each state. Thus for next step I will put state as a radom effecto for the multilevel generalized linear model.

4. Model Analysis

4.1 Multilevel linear model for option C

First I want to find out the relationship between the option C each customer finally chose with the option C they chosen before.

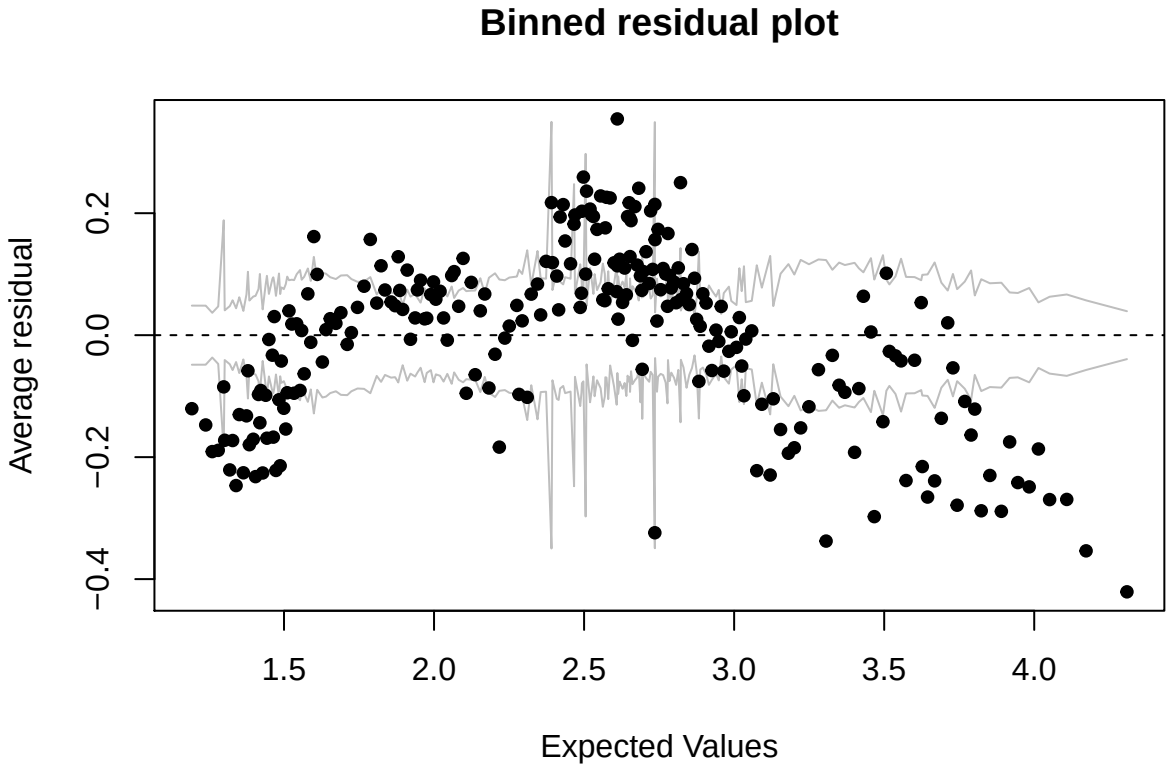
$C_i = Group\_Size_{j[i]} + State_{k[i]} + Car\_Value_i + Risk\_Factor_i + Married\_Couple_i + C\_Previous_i + \epsilon_i$
$C_i \sim N(\mu_C, \sigma_C^2)$ $\epsilon_i \sim N(0, \sigma_C^2)$ $Group\_Size_j \sim N(\mu_{Group\_Size}, \sigma_{Group\_Size}^2)$ $State_k \sim N(\mu_{State}, \sigma_{State}^2)$

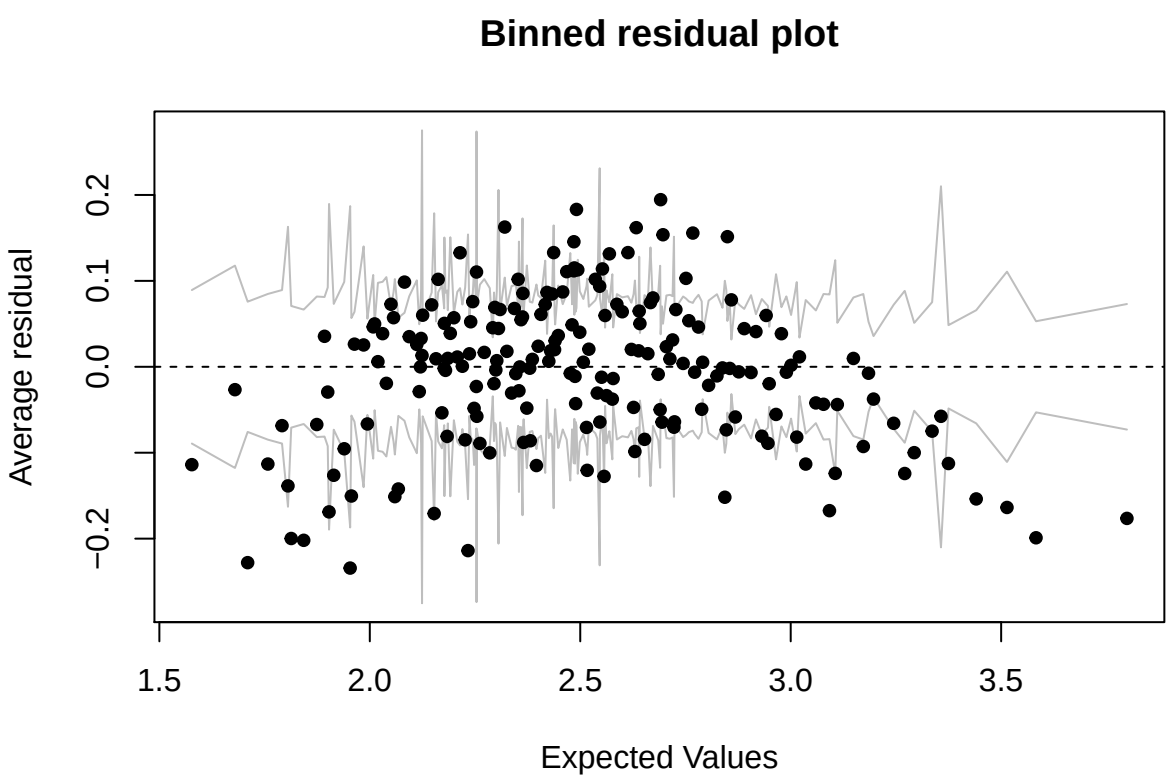
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: C ~ car_value + risk_factor + C_previous + married_couple + (1 |
##     state) + (1 | group_size)
##    Data: purchase
##
##      AIC      BIC   logLik deviance df.resid
## 182727.0 182853.6 -91349.5 182699.0    62241
##
## Scaled residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.57246 -0.30041  0.06735  0.24105  2.68967
##
## Random effects:
##  Groups     Name        Variance  Std.Dev.
##  state      (Intercept) 0.0057137 0.075589
##  group_size (Intercept) 0.0000943 0.009711
## Number of obs: 62255, groups:  state, 36; group_size, 4
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.056412   0.071116    0.79 0.427642
## car_valueb      0.123871   0.095036    1.30 0.192434
## car_valuec      0.130924   0.070326    1.86 0.062648 .
## car_valued      0.168497   0.068767    2.45 0.014276 *
## car_valuee      0.199691   0.068595    2.91 0.003601 **
## car_valuef      0.230446   0.068608    3.36 0.000783 ***
## car_valueg      0.253094   0.068716    3.68 0.000230 ***
## car_valueh      0.266913   0.069298    3.85 0.000117 ***
## car_valuei      0.287615   0.074117    3.88 0.000104 ***
## risk_factor    -0.046852   0.002540  -18.44  < 2e-16 ***
## C_previous      0.281705   0.002859   98.52  < 2e-16 ***
## married_couple  0.041480   0.010488    3.95 7.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) cr_vlb cr_vlc cr_vld car_valuee cr_vlf cr_vlg cr_vlh
## car_valueb  -0.689
## car_valuec  -0.937  0.696
## car_valued  -0.958  0.712  0.969
## car_valuee  -0.960  0.714  0.971  0.993
## car_valuef  -0.959  0.714  0.971  0.993  0.995
## car_valueg  -0.957  0.713  0.969  0.991  0.994      0.993
## car_valueh  -0.948  0.706  0.961  0.983  0.985      0.985  0.984
## car_valuei  -0.885  0.660  0.898  0.919  0.921      0.921  0.919  0.912
## risk_factor -0.115 -0.005 -0.003 -0.004 -0.004     -0.004 -0.004 -0.003
## C_previous  -0.135  0.008  0.014  0.011  0.007      0.004 -0.001 -0.005
## married_cpl -0.070  0.004  0.006  0.005  0.004      0.000 -0.001 -0.002
##             car_valuei rsk_fc C_prvs
## car_valueb
## car_valuec
## car_valued
## car_valuee
## car_valuef
## car_valueg
## car_valueh
## car_valuei
## risk_factor -0.007
## C_previous  -0.013      0.120
## married_cpl  0.004      0.037  0.014
## convergence code: 0
## Model failed to converge with max|grad| = 0.0011903 (tol = 0.001, component 1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: C ~ car_value + risk_factor + married_couple + (1 | state) +
##     (1 | group_size)
##    Data: purchase
##
##      AIC      BIC   logLik deviance df.resid
## 192924.5 193042.0 -96449.3 192898.5    62242
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.41107 -0.46233  0.08336  0.41941  2.10409
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  state      (Intercept) 0.019404 0.13930
##  group_size (Intercept) 0.001566 0.03957
## Number of obs: 62255, groups:  state, 36; group_size, 4
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.993484   0.077126  12.881  < 2e-16 ***
## car_valueb      0.040421   0.096042   0.421  0.67385
## car_valuec      0.032654   0.071057   0.460  0.64584
## car_valued      0.091685   0.069513   1.319  0.18718
## car_valuee      0.152291   0.069346   2.196  0.02808 *
## car_valuef      0.207038   0.069359   2.985  0.00284 **
## car_valueg      0.262291   0.069471   3.776  0.00016 ***
## car_valueh      0.305334   0.070052   4.359 1.31e-05 ***
## car_valuei      0.386519   0.074813   5.166 2.39e-07 ***
## risk_factor    -0.079218   0.002523 -31.403  < 2e-16 ***
## married_couple  0.029642   0.009907   2.992  0.00277 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) cr_vlb cr_vlc cr_vld car_valuee cr_vlf cr_vlg cr_vlh
## car_valueb  -0.646
## car_valuec  -0.872  0.702
## car_valued  -0.891  0.718  0.969
## car_valuee  -0.893  0.719  0.972  0.993
## car_valuef  -0.893  0.719  0.971  0.993  0.995
## car_valueg  -0.891  0.718  0.970  0.991  0.994      0.994
## car_valueh  -0.884  0.712  0.962  0.983  0.986      0.985  0.984
## car_valuei  -0.827  0.667  0.900  0.920  0.923      0.923  0.921  0.914
## risk_factor -0.094 -0.004 -0.003 -0.003 -0.003     -0.003 -0.003 -0.001
## married_cpl -0.091  0.003  0.008  0.007  0.005      0.001  0.000 -0.002
##             car_valuei rsk_fc
## car_valueb
## car_valuec
## car_valued
## car_valuee
## car_valuef
```

```
## car_valueg
## car_valueh
## car_valuei
## risk_factor -0.004
## married_cpl  0.004      0.035
```






```
## Data: purchase
## Models:
## c_purchase2: C ~ car_value + risk_factor + married_couple + (1 | state) +
## c_purchase2:     (1 | group_size)
## c_purchase1: C ~ car_value + risk_factor + C_previous + married_couple + (1 |
## c_purchase1:     state) + (1 | group_size)
##             Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## c_purchase2 13 192925 193042 -96449   192899
## c_purchase1 14 182727 182854 -91350   182699 10199      1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. The AIC is lower while I add C_previous to the model, this prove that the previous option on C will influence the final purchase on option C.
2. The variance of random effect (predictor state and group_size) in model 2 is larger than model 1, this shows that the random effect has more influence to this model.

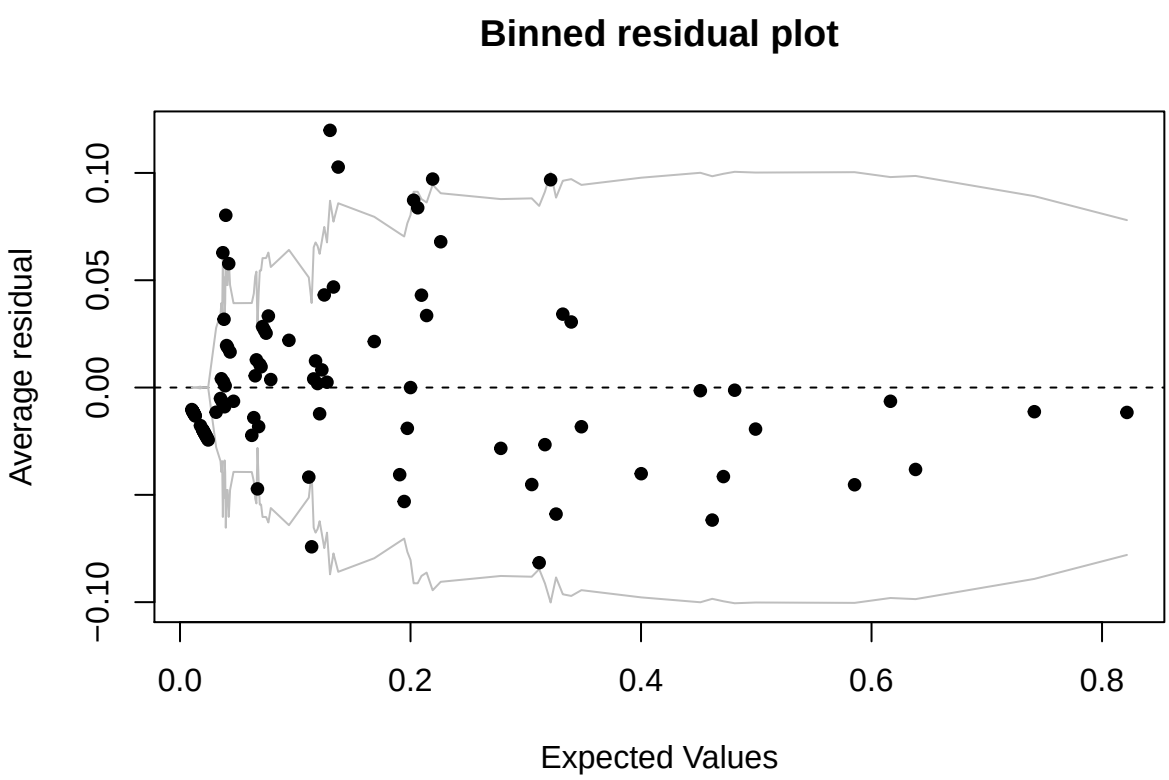
4.2 Logistical model for purchase prediction with previous c option

Next I will predict the combination of the options by fit the data into multilevel logistic regression model. I will use two partial pooling model, one is random intercept, the other is random intercept and random slope.

$Prob(record\_type = 1) = logit^{-1}(Plan_{j[i]} + Duration\_Previous_i + shopping\_point_i + C\_Previous_i + \epsilon_i)$
$\epsilon_i \sim N(0, \sigma_C^2)$ $Plan_j \sim N(\mu_{Plan}, \sigma_{Plan}^2)$

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 10) [glmerMod]
```

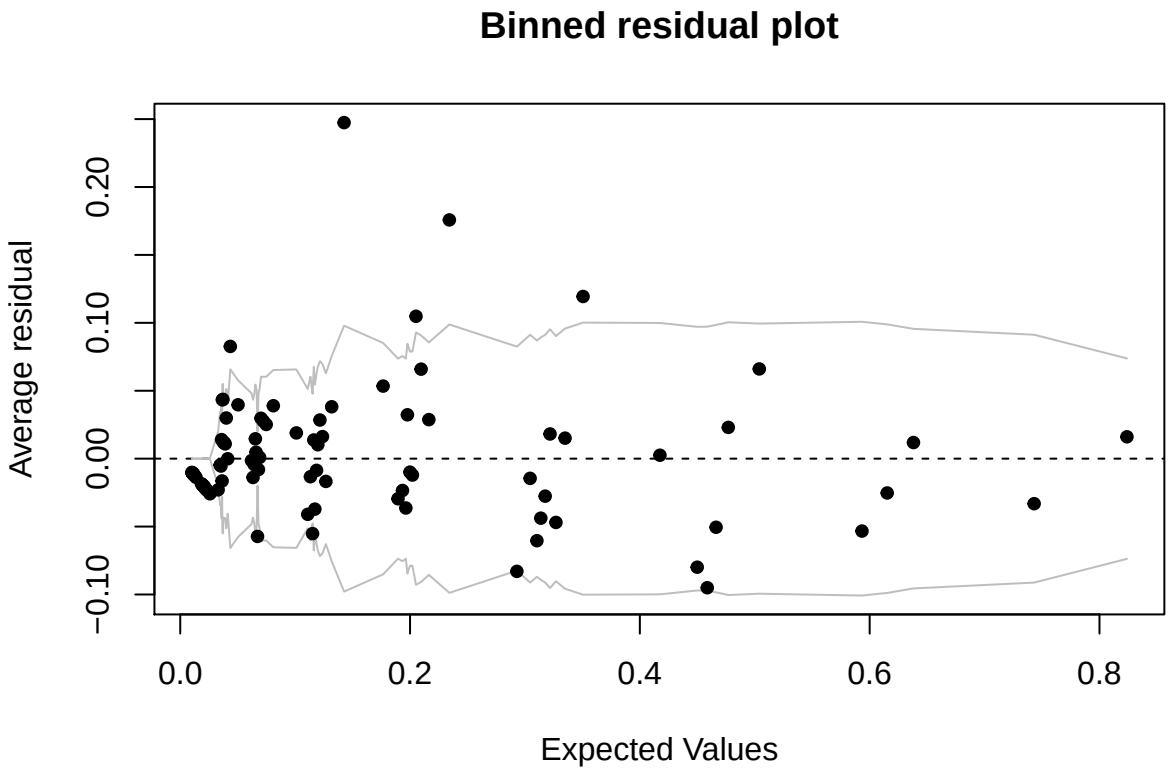
```
##  Family: binomial  ( logit )
## Formula: record_type ~ C_previous + duration_previous + shopping_pt +
##     (1 | plan)
##    Data: test
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##   6261.2   6297.3  -3125.6   6251.2     9995
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3607 -0.3718 -0.2005 -0.1110  5.3466
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  plan   (Intercept) 0.01583  0.1258
## Number of obs: 10000, groups:  plan, 947
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -5.22563    0.14498  -36.04   <2e-16 ***
## C_previous         0.02792    0.03264    0.86   0.3923
## duration_previous  0.01303    0.00681    1.91   0.0556 .
## shopping_pt        0.61826    0.01651   37.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) C_prvs drtn_p
## C_previous  -0.556
## duratn_prvs -0.288 -0.105
## shopping_pt -0.733  0.009  0.039
```



	Est	LL	UL
(Intercept)	-5.2256304	-5.5097888	-4.9414719
C_previous	0.0279222	-0.0360592	0.0919035
duration_previous	0.0130354	-0.0003113	0.0263820
shopping_pt	0.6182635	0.5858965	0.6506306

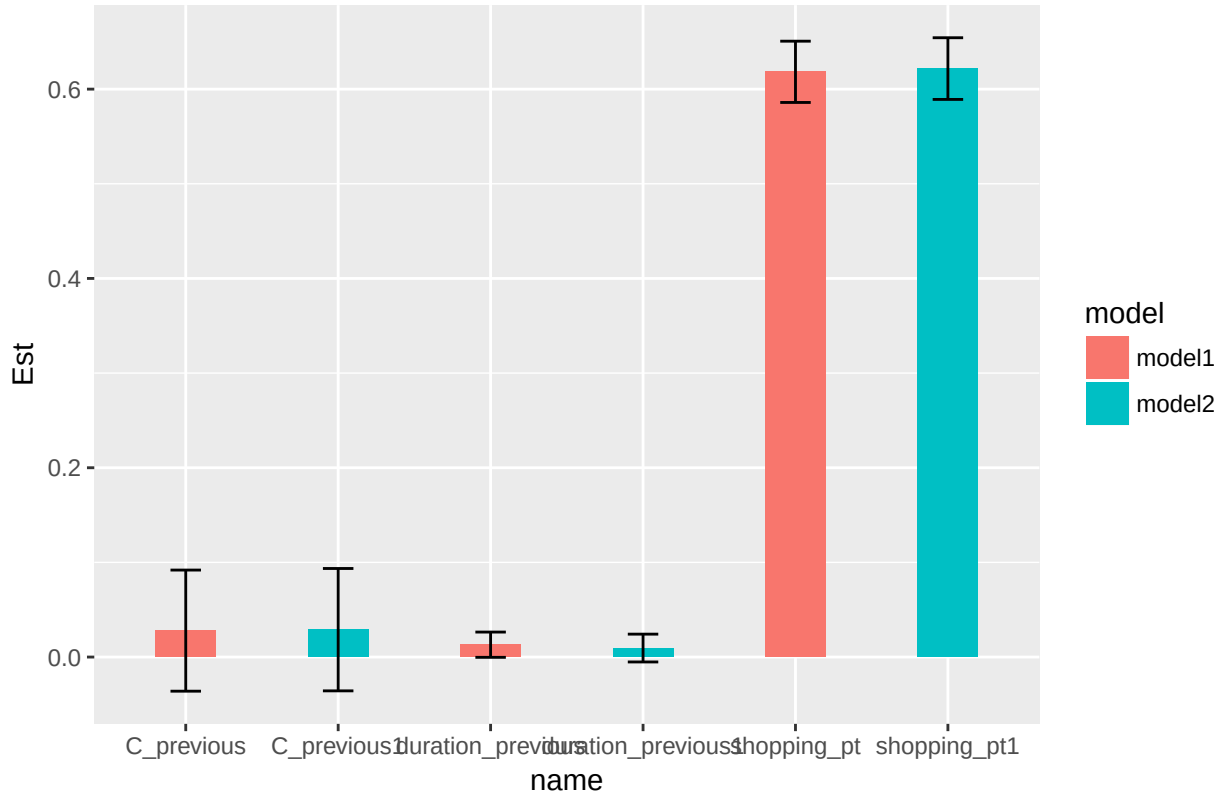
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: record_type ~ C_previous + duration_previous + shopping_pt +
##     (1 + duration_previous | plan)
##    Data: test
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##   6262.5   6313.0  -3124.2   6248.5     9993
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3039 -0.3670 -0.1976 -0.1086  5.2948
##
## Random effects:
##  Groups Name              Variance Std.Dev. Corr
##  plan   (Intercept)       0.001131 0.03363
##         duration_previous 0.000900 0.03000  -1.00
```

```
## Number of obs: 10000, groups:  plan, 947
##
## Fixed effects:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -5.232601   0.143898  -36.36   <2e-16 ***
## C_previous         0.028917   0.032982    0.88    0.381
## duration_previous  0.009507   0.007500    1.27    0.205
## shopping_pt        0.621699   0.016621   37.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) C_prvs drtn_p
## C_previous  -0.569
## duratn_prvs -0.256 -0.088
## shopping_pt -0.725  0.011 -0.010
```



	Est	LL	UL
(Intercept)	-5.2326015	-5.5146424	-4.9505606
C_previous	0.0289171	-0.0357285	0.0935626
duration_previous	0.0095067	-0.0051934	0.0242068
shopping_pt	0.6216994	0.5891221	0.6542767

Figure 4.1 Means and error bars for coefficient




```
## Data: test
## Models:
## predict.1: record_type ~ C_previous + duration_previous + shopping_pt +
## predict.1:     (1 | plan)
## predict.2: record_type ~ C_previous + duration_previous + shopping_pt +
## predict.2:     (1 + duration_previous | plan)
##           Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## predict.1  5 6261.2 6297.3 -3125.6   6251.2
## predict.2  7 6262.5 6313.0 -3124.2   6248.5 2.7425      2     0.2538
```

From the figure 4.1 we can see that two model and similar standard error and there is not much difference between the two models with random intercept and random intercept + random slope. The conclusion is also proved by the anova, thus I will choose predict.1 as my model.

5. Conclusion

Comment

This dataset is not complete and lack the data for those customors who end up without purchase any insurance, thus I can not fit an accurate model and make prediction from it.