Funeral Planning

Jennifer Mo, Andrew Barton, Yue Ren; TF mentor: Jun Li; Professor Mentor: Masanao Yajima

12/17/2016

1. Background and Data Format:

1.1 Overview

In this project, our client Sarah Whitney from Department of Marketing, Questrom School of Business is interested in whether who plans the funeral (the self versus another) impacts spending on the funeral, taking into consideration all the different cost categories.

Our client has came to our consulting services in the summer and got advice about modeling, now our client is focusing on two-stage model:

(1) Binary(expense > 0 or not) ~ category + self indicator + private indicator + self:private interaction + (1|Funeral ID)

(2) Positive expense only ~ category + self indicator + private indicator + self:private interaction + (1|Funeral ID)

We have got the data for the second model, that is, the entries where the expenses are positive.

Our client has two concrete questions:

(1) Dealing with the 2 models, how should we relate the two together for interpretation

(2) How to interpret the interaction term (in SPSS)?

After looking into our client's SPSS result, we found some coefficients estimations have very large standard error. Now we have the third question:

(3) Correct the standard error explosion problem.

1.2 Problem Solving Strategy and Data Cleaning

Question (3) needs to be solved first. After investigation, we found that: if we put more than two of these three predictors in the model: Expense_type, category, and public (all of them are factors), there will be collinearity issue. A particular category is a combination of several expense_type. Category 1 is public while Category 2 are private. (Only Category 1 and 2 should be included in the analysis.) We have to include at most one of these variables in the model.

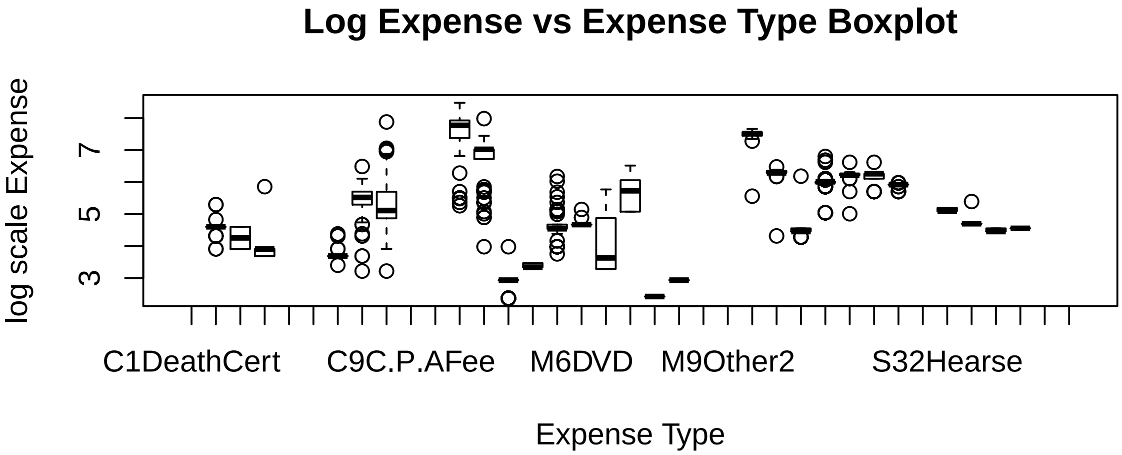
After data cleaning procedure, the data format is like this:

```
head(funeral_sub)
```

```
##     Expense Expense_Binary Category Expense_Type Other_Planned Public ID
## 362     100              1        2     C2Clergy             0      0 11
## 363     100              1        2     C2Clergy             0      0 24
## 364     100              1        2     C2Clergy             0      0 27
## 365     100              1        2     C2Clergy             0      0 30
## 366      75              1        2     C2Clergy             0      0 37
## 367     100              1        2     C2Clergy             0      0 46
```

2. Check variability and Model Selection

There is very large variance in the range of costs between the different types of expenses, but because the collinearity issue when treating it as fixed effect, we could only take it as random effect, in additional to ID.



The y axis is the log scale of expense. Log scale could shrink some variability, it is also widely used in modeling cost or revenue. Because the Expense_type's and ID's are not nested, we used the following model:

```
fit1<-lmer(log(Expense) ~ Other_Planned + Public + Other_Planned:Public +
                          (1|ID) + (1|Expense_Type), data=funeral_sub)
round(summary(fit1)$coefficient ,2)
```

```
##                      Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)              4.55       0.29   25.26   15.52     0.00
## Other_Planned            0.00       0.01  258.50    0.14     0.89
## Public                   1.17       0.51   25.13    2.30     0.03
## Other_Planned:Public     0.04       0.02 3646.92    2.25     0.02
```

Which model should we used? Let's do some model comparison:

```
fit2 <- lmer(log(Expense) ~ Other_Planned + Public + (1|ID) + (1|Expense_Type), data=funeral_sub)
fit3 <- lmer(log(Expense) ~ Other_Planned  + (1|ID) + (1|Expense_Type), data=funeral_sub)
anova(fit1, fit2)
anova(fit2, fit3)
```

I omitted the results here but the ANOVA shows fit1 is the best model among them.

3. Z score Usage Discussion

When discussing this project with our client, we have talked about if we should z-score the expense/log scale expense within expense type. After discussing it with our professor, we have several reasons not use this:

(1) If we z-scored a vector of expenses by substracting the mean and dividing them using the standard error, even if there exists effect of public or private(yes, there exists, see the result in Section 2), it will be cancelled. So firstly, never z-scored expenses within expense type in this way.

(2) If we z-scored a vector of expense by dividing them using the standard error without substracting the mean, it conflicts our data natural. It is common that a more expensive type have larger variability. This procedure will cause some unpredictable result. I tried it in the r file and you will see the significance of public is cancelled again.

(3) The variance of expense or log scale expense comes from several source, z-scored the data via method (1) or (2) just simple made the variance of expense in each expense type the same. Using this method will lose a lot of information. The Significance lose in (1) and (2) shows this effect.

4. Answering Concrete Questions Formally:

(1) Dealing with the 2 models, how should we relate the two together for interpretation?

Answer: The first model is related to the probability that the expense is positive. The second model is about: conditional on the expense is positive, how does the expense or the log scale of the expense relate to the predictors. That is the link between the two models.

Here is an example, given the funeral is Self_Planned, the probability of the expense change from private to public is explained by model (1). Still fix the Self_Planned to be true, condition on the expense is positive, the expectation of Public expense is `exp(1.17)=3.222` times of the expectation of Private expense. We should also consider the coefficients of public in model (1) to determine if the predictor public makes a large difference.

(2) How to interpret the interaction term (in SPSS)?

Answer: We agreed that we ran the models in R and our client could compare results between R and SPSS. Now looking at the output in Section 2, we will found the baseline is self planned and private. Here are some examples of how the interactions comes into interpretation, conditional on the expense is positive:

(i) fix self-planned unchanged, Expense(public) = exp(1.17)*Expense(private) on average
(ii) fix public type to be private, Expense(Other planned)) = exp(0)*Expense(Self planned) = Expense(Self planned) on average
(iii) fix other-planned unchanged, Expense(public) = exp(1.17+0.04)*Expense(private) on average. Here is where the interaction term appears
(iv) fix public type to be public, Expense(Other planned) = exp(0+0.04)*Expense(Self Planned) on average

(3) Correct the standard error explosion problem.

Answer: Corrected it in Section 2.