Travel Demand Analysis
Exercise 2

310702051 葉家榮

2022-03-22

This assignment is also uploaded in the following website, for detailed information and code. https://chiajung-yeh.github.io/Travel-Demand-Analysis/logistic-poisson-regression.html

Problem

1. Use simulation to demonstrate the following collider bias example (see the figure below).
   Step 1: Generate three random variables, X, Y, and Z, all following a binomial distribution. Among them, X and Y are independently generated while Z is generated depended on X and Y.
   Step 2: Run a logistic regression of Y on X (there should be no relationship).
   Step 3: Run a logistic regression of Y on X and Z (there should be some relationship).
   Note: *binomial distributions are a convenient choice; you can use other distributions if you like.*

2. Based on the National Household Travel Survey data of Arizona State, answer the following questions.

   (1) Use household vehicle ownership (HHVEHCNT) as the outcome variable. Run a single variable Poisson regression with either household size (HHSIZE), number of adults in the household (NUMADLT), or number of drivers in the household (DRVRCNT) as the predictor. Compare these three models, especially the estimated coefficients and model fit.

   (2) Develop a logistic regression to predict "fully-equipped household" (You need to mutate a binary variable based on "vehown.fctr"). You need to incorporate at least one household variable, one built-environment variable, and on interaction term. Explain the estimation result.

Collider Bias

Collider bias is used to describe the variable which is simultaneously influenced by the independent variable and the dependent variable in the model formulation. Though the causal variables influencing the collider are themselves not necessarily associated, that is, the dependent variable and the independent variable of the model have no casual effect, the spurious relationship is found under the intervention of collider variable. The casual graph of collider bias is shown in Figure 1.

The simulation is conducted below to demonstrate the collider bias. Let the independent variable (X) and dependent variable (Y) be binomial random variable, and let the collider variable (Z) also be under binomial distribution, but associated with X and Y. Here we force the probability of binomial distribution to be the inverse logit of X plus Y. The code is shown below, and the relationship of three variables are shown in Figure 2. We can find that the frequency of X and Y are approximately the same, while X and Z, Y and Z are apparently not independent. We then take Y as dependent variable, and construct two logistic regression models. For the first model, X is the only independent variable, while X and Z are both independent variable in the second model.

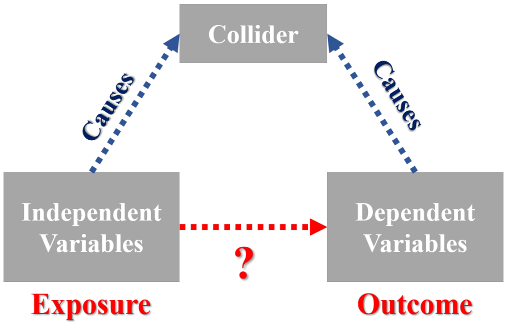

Figure 1: Casual graph of collider bias

```r
N=1000
set.seed(999)

# set the binomial random variable
X=rbinom(N, 1, 0.5)
Y=rbinom(N, 1, 0.5)
Z=rbinom(N, 1, invlogit(X+Y))

# logistic regression (Y~X)
glm(Y ~ X, family=binomial("logit"))

# logistic regression (Y~X+Z)
glm(Y ~ X+Z, family=binomial("logit"))
```

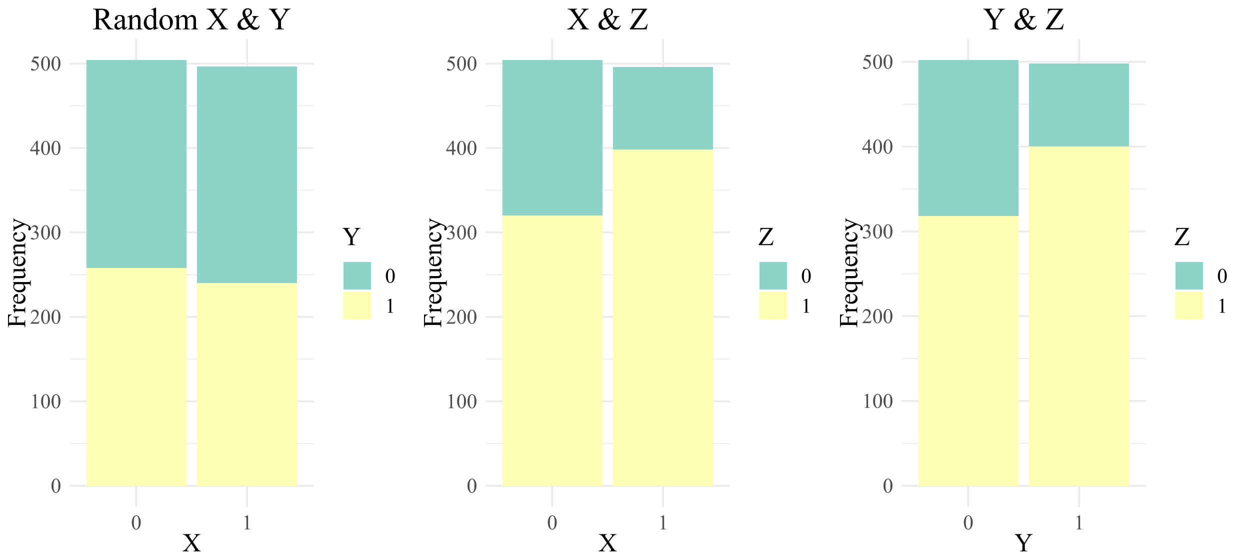

Figure 2: Relationship between X, Y, and Z

The result of logistic regression is shown in Table 1. We can find that X and Y has no significant relationship in ***Original Model***. It is reasonable, since both X and Y are both random variable, they have no association. But interestingly, X and Y has a significant relationship in ***Collider Bias Model***, it is because variable Z is included in the model, and cause a spurious relationship. The phenomenon illustrated by this simulation is called “Collider Bias”. It suggests that collider variable should be removed to avoid the fallacy.

Table 1: Collider bias model result

	Dependent variable:	
	Y	
	Original Model	Collider Bias Model
	(1)	(2)
X	-0.112	-0.273^{**}
	(0.127)	(0.132)
Z		0.919^{***}
		(0.149)
Constant	0.048	-0.538^{***}
	(0.089)	(0.133)
Observations	1,000	1,000
Log Likelihood	-692.746	-672.984
Akaike Inf. Crit.	1,389.492	1,351.967
Note:		*p<0.1; **p<0.05; ***p<0.01

Take practical transport issue for instance. If we want to measure the effect of the private vehicles restriction on the modal share of public transport, the independent variable would be “private vehicles restriction”, while the dependent variable would be “modal share of public transport” in the model. Here, we add an independent variable reflecting the “air pollution”, it might cause the model to be biased since the two variables are much correlated to “air pollution”. Hence, in this model, “air pollution” can be termed as collider variable, and it is better to remove the collider one.

Poisson regression

Based on the National Household Travel Survey data of Arizona State, take household vehicle ownership variable as the outcome variable, and let the household size, number of adults, number of drivers be the independent variable respectively. The model results are shown in Table 2.

The coefficient of HHSIZE in Model 1 is the expected difference in household vehicles (on the logarithmic scale) for each additional household members. Thus, the expected multiplicative increase is $e^{0.138} = 1.148$, namely, a 14.8% positive difference in the household vehicles per household member. Similarly, in Model 2, if the number of adults increases by one, the number of household vehicles is expected to increase $e^{0.322} - 1 =$ 38.0%. Last, in Model 3, if the number of drivers increases by one, the number of household vehicles would increase $e^{0.389} - 1 = 47.6\%$.

As for the model fit of each three models, we can find that the log likelihood is the largest in Model 3, representing that Model 3 offers a better fit to the data compared to others. Note that the log likelihood value of null model which only contains the constant term is -5095.911 ($LL(0)$), while the log likelihood is -3647.652 ($LL(\beta)$) for Model 3. It implies that the model indeed improves a lot compared to the null model. A formal method to examine the model is to conduct the likelihood ratio test. The Chi-square of the two model is 2896.5 (p-value is very close to 0), and hence, we can say that Model 3 is better than the null model, and best fit among the three. In addition, AIC and deviance also provide the model fit information, the lower the better. The prediction curve and the real data is shown in Figure 3.

Using number of drivers as the dependent variable is the best one to predict the household vehicle ownership. It is because that number of drivers is more correlated to the outcome variable (vehicle ownership), they are the most potential users of the vehicle, and probably own a car. As for the number of adults, not all the adults have driver license, and hence not much correlated to vehicle ownership relative to number of drivers.

Table 2: Poisson regression on HHVEHCNT

	Dependent variable:		
	HHVEHCNT		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
HHSIZE	0.138***		
	(0.010)		
NUMADLT		0.322***	
		(0.018)	
DRVRCNT			0.389***
			(0.019)
Constant	0.471***	0.163***	0.058
	(0.028)	(0.038)	(0.039)
Observations	2,414	2,414	2,414
Log Likelihood	−3,770.225	−3,697.150	−3,647.652
Akaike Inf. Crit.	7,544.450	7,398.301	7,299.304
Note:			*p<0.1; **p<0.05; ***p<0.01

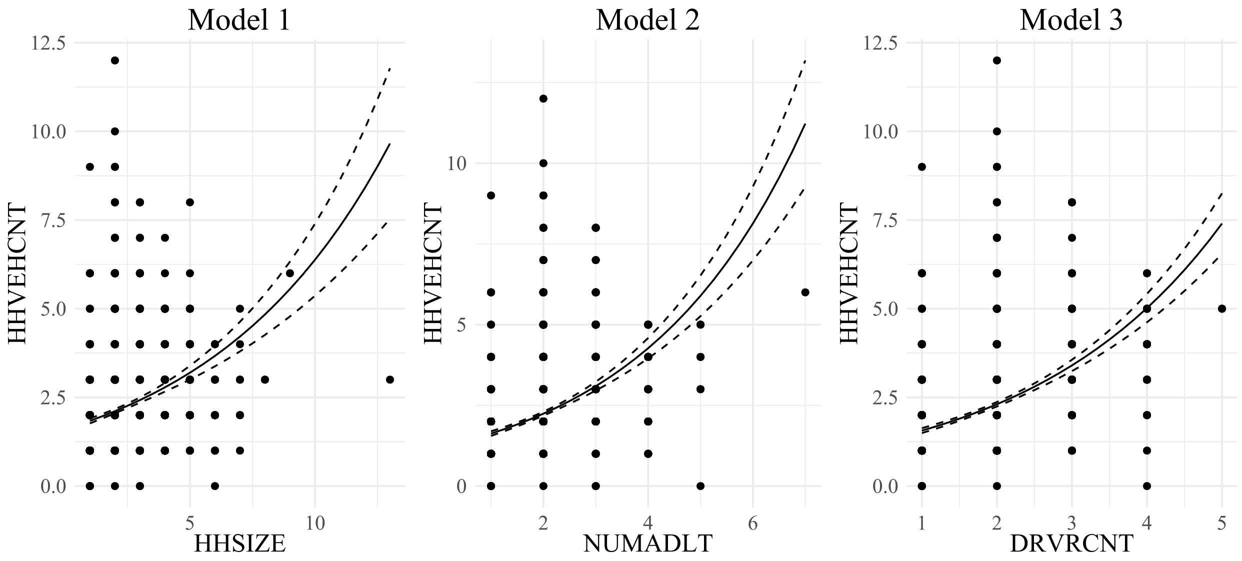

Figure 3: Model fit visualization

Last, household size also has a significant impact on the vehicle ownership though, it has the lowest effect. A possible reason is that household may consist of kids, who are absolutely prohibited to have a driver license, and thus have no vehicle usually. In conclusion, number of drivers can explain the most variance to predict the household vehicle ownership, for it is highly associated with vehicle ownership.

Logistic regression

To predict "fully-equipped household" by developing the logistic regression, we mutate a binary variable based on "vehown.fctr" in the data. Before formulating the model, it is vital to clearly understand the definition of "fully-equipped household". The term means whether there are at least 1 car per driver in the household, and it is calculated by HHVEHCNT and DRVRCNT ($HHVEHCNT/DRVRCNT$). With this background knowledge, it is suggested not add two variables simultaneously in the model, or it might make the logistic regression nonidentified. Let us first show the model by considering both two variables, illustrated in Table 3.

```
glm(FullyEquip ~ HHVEHCNT+DRVRCNT, family=binomial("logit"), data=dat_AZ)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Table 3: Nonidentified logistic regression model result

	Dependent variable:
	FullyEquip
HHVEHCNT	48.601
	(7,543.801)
DRVRCNT	−49.080
	(8,701.359)
Constant	25.796
	(10,654.150)
Observations	2,414
Log Likelihood	−0.00000
Akaike Inf. Crit.	6.000
Note:	*p<0.1; **p<0.05; ***p<0.01

In the model result, we can first find the "Warning", which shows that "algorithm did not converge", also the standard error of each variables and constant are abnormally high. This problem will occur if any linear combination of predicton is perfectly aligned with the data, with $y = 1$ if and only if this linear combination of predictors exceeds some threshold. In this case $HHVEHCNT/DRVRCNT$ exactly determines the threshold of "fully-equipped household", if the value is not less than one, it is classified as "fully-equipped household", while not if the value is less than 1. And thus, placing two key factors in the model would directly decide whether fully-equipped or not. The two factors are not allowed to put in a same logistic regression model.

Now, take one household variable (standardization of household size), one built-environment variable (whether household live in urban or not), and the interaction term (HHSIZE:urban) to develop a new logistic regression. Note that the correlation of HHSIZE and urban is close to 0, indicating that the variable is reasonable to place in the same model. The model result is shown in Table 4.

Table 4: FullyEquip logistic regression model result

	Dependent variable:
	FullyEquip
c.hhsize	-0.257^{**}
	(0.104)
factor(urban)1	-0.159
	(0.191)
c.hhsize:factor(urban)1	-0.115
	(0.120)
Constant	2.497^{***}
	(0.172)
Observations	2,414
Log Likelihood	-715.577
Akaike Inf. Crit.	1,439.154
Note:	*p<0.1; **p<0.05; ***p<0.01

In Table 4, we can find that the coefficient of standardization of household size (c.hhsize) is -0.257, meaning that add one standard deviation of household size would drop $e^{-0.257} = 0.77$ times of odds of fully-equipped. If the household lives in urban area, the odds of fully-equipped would decrease $e^{-0.159} = 0.85$ times. Though this variable is not significant, the sign is expected. Since people would tend to have no car in the urban area, and use public transport instead (this is true for most of Asian cities, but I have no idea whether it is also true in Arizona), the negative sign proves that the household has a lower odds to be fully-equipped compared to the non-urban area. According to Chapter 4.5 in *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Gelman list the general principles of building regression models, and he suggests that if a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. Thus, retain the urban variable in the model might be reasonable. Also, the interaction variable in the model, we can find that the sign is also negative, which shows that if the household size is fixed, household living in urban would have a more negative effect on the odds of fully-equipped. Again, this inference might be weak, due to the coefficient of interaction term is also insignificant.