Factorial Design for Factors Affecting Total Tourist Expenditure

STA305 H1S Sec L0201, W2022

Yutong Lu 1005738356

2022-04-08

Introduction

With the lifting of COVID-19 restriction in many areas, tourism is expected to recover from the effects brought by the global pandemic, including the restrictions on hotels, travelling, restaurants and recreation. According to Marrocu, Paci and Zara (2015), tourist income, foreign nationality and employment status have significant effects on the tourist expenditure. Another study agreed with the effect of country of origin but also found trip length, family travelling, first visit and activity participation also have prominent roles on the total expenditure (Almeida & Garrod, 2017). However, Vieira and Santos (2012) did not find the effect of length of stay significant, rather they concluded that the number of areas visited has a greater influence on the money spent, which may correspond to the effect of pursued activities.

Therefore, the research question is whether having a foreign nationality, travelling with family, first visit and the activity level of the visit can influence the total tourist expenditure. Based on the literature, we hypothesized that being a foreign tourist, travelling with family, first visit of the area and being an active visitor have important impacts on total expenditure. We aimed to investigate the main effects and interactions of nationality, travelling with family, first visit and activity level on the total tourist expenditure. This study is important because we can provide tourists, business owners, and policymakers with a clearer idea about the important factors on tourist expenditure and help them improve their experience and business as we are moving towards a post-pandemic world.

Materials and Methods

Experimental Design and Data The four factors in this design are foreign nationality, travelling with family, first visit and activity level, each with two levels corresponding to true and false, or low and high. The response in this study is the total tourist expenditure, measured in CAD. Because some areas have higher or lower prices due to regional differences, there are also two blocks in this design. Because the objective was to identify significant main effects and interactions, we constructed a 2^4 factorial design with two blocks of size 8, confounded with higher-order interactions. Because there was no prior estimate of error and no available evidence supporting that some interactions are negligible, the design was replicated five times to obtain an estimate of error. We also used a partial confounded approach (Yates, 1937), where in each of the five replicates, the blocking was confounded with a different interaction. Note that no main effects were used for confounded with blocks.

We denoted having a foreign nationality as factor A, travelling with family as factor B, first visit as factor C and the activity level as factor D. Each factor A, B, C, and D had two levels, -1 and 1. For factor A, level -1 represented the tourist stayed within the country of residence, whereas level 1 represented that the tourist was from a foreign country. Level -1 of factor B represented the tourist was not travelling with family, and level 1 represented the tourist was travelling with family members. Factor C level -1 represented the tourist who had visited the area before, and level 1 represented that the tourist was a first-time traveller to this place. For factor D, level -1 represented the tourist was not active, whereas level 1 represented the tourist was actively pursuing activities during the trip. The five replicates were denoted as replicate I, II, III, IV and V.

Higher-order interactions are usually deemed as either negligible or less important relative to the lower-order interactions and main effects (Montgomery, 2020). Thus, in this experimental design, we chose the highest two orders of interactions ABCD, ABC, ABD, ACD and BCD to be confounded with blocks in the five replicates I, II, III, IV, and V, respectively. Under this design, the information of ABC could be obtained

from replicates I, III, IV, and V; ABD interaction information can be obtained from replicates I, II, IV, and V; ACD information could be obtained from I, II, III, and V; BCD information could be obtained from replicates I, II, III, and IV; and the information of interaction ABCD can be obtained from replicates II, III, IV, and V. Thus, we created the column for blocking (either Block 1 or Block 2) using a defining contrast

$$L = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4$$

where $x_i = 0$ indicates the low level of the i-th factor and $x_i = 1$ indicates the high level of the i-th factor. $\alpha_i = 0,\ 1$ for $i = 1, 2, 3, 4$ is the exponent of x_i in a certain combination of $A^{\alpha_1} B^{\alpha_2} C^{\alpha_3} D^{\alpha_4}$. For the five replicates, we used five different defining contrasts depending on the confounding interaction. For example, in replicate II where ABC was confounded with blocks, $L = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$. Then, we calculated L for each combination and evaluated $L \ mod\ 2$. Combinations with $L \ mod\ 2 = 0$ were set to be in Block 1 and those with $L \ mod\ 2 = 1$ were set to be in Block 2.

We then simulated data for the design. Based on the literature, foreign tourists, travelling with family, first-time visitors and high activity levels all can positively affect the total expenditure (Marrocu, Paci & Zara, 2015; see also Almeida & Garrod, 2017; Vieira & Santos, 2012). As a result, we arbitrarily inputted values for the grand mean and each of the main effects and interactions based on literature evidence (*Supplementary Table 1*). As suggested in the literature, any interactions with the factor A representing foreign travelers were inputted with greater values for effect. Also, second and third-order interactions were inputted with smaller effect values because they were considered as less important (Montgomery, 2020). We also randomly generalized errors from a Normal distribution with a mean of 0 and a standard deviation of 100. The response for each of the combinations in the four replicates was then generated by adding up the corresponding main effects, interactions and the random error.

Statistical analysis With the simulated data, we fitted a linear regression model using the total tourist expenditure as the response, the four factors A, B, C and D for main effects, interactions and block effects as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_{1234} x_1 x_2 x_3 x_4$$

where $x_i,\ i = 1, .., 4$ indicates the four factors A, B, C, and D. $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1, ..., \hat{\beta}_{1234}$ are the coefficient estimates. Then, we would check the estimated coefficients and calculate the factorial effect estimates by multiplying the corresponding coefficient estimates by 2. Note that the estimated intercept $\hat{\beta}_0$ itself is the estimated grand mean $\hat{\mu}$. However, because we arbitrarily inputted values for main effects and interactions, the focus of our study was not on the effects estimates but rather on the analysis of variance.

Therefore, we checked the source of variance and the degrees of freedom using the ANOVA table. We partitioned the total degrees of freedom to replicates, blocks within replicates, main effects and interactions, and error. Then, we calculated the sum of squares of replicates using $SS_{rep} = \sum_{h=1}^{n} \frac{R_h^2}{2^k} - \frac{y_{...}^2}{N}$ where R_h is the sum of response for the hth replicate, k is the number of factors, $y_{...}^2$ is the squared sum of all responses and N is the total number of observations. We also calculated sum of squares for blocks and error using $SS_{block} = SS_{ABCD,\ repI} + SS_{ABC,\ repII} + SS_{ABD,\ repIII} + SS_{ACD,\ repIV} + SS_{BCD,\ repV}$ and $SS_{error} = SS_{total} - SS_{rep} - SS_{block} - SS_A - ... - SS_{ABCD} - SS_{error}$. Note that the total sum of square was obtained by summing up all the sums of squares in the ANOVA table.

According to Montgomery (2020), normal or half-normal plots and Lenth's method are analysis procedures for unreplicated two-level factorial design. Because there were five replicates in our study, it was not appropriate to use normal plots or Lenth plots to identify active main effects and interactions. As a result, we would only find the significant main effects and interactions by checking the p-values in the ANOVA table and identifying any significant terms with p-values smaller than a significance level of 0.05. We would then fit these significant effects and interactions into a new linear regression model and check ANOVA table again to see whether the main effects and interactions are still important. Finally, for both the initial model and the new model, we checked the model assumptions of heterogeneity of error variance and error normality using residual plot and Normal quantile-quantile plot. If the assumptions were satisfied, we would expect no patterns in the residual plot and the point in the Normal quantile-quantile plot to align closely with the diagonal line within the boundaries.

Table 1: Partially Confounded 2^4 Design with 5 Replicates.

					Replicate I ABCD Confounded	Replicate II ABC Confounded	Replicate III ABD Confounded	Replicate IV ACD Confounded	Replicate V BCD Confounded
Combination	A	B	C	D	Block	Block	Block	Block	Block
(1)	-1	-1	-1	-1	1	1	1	1	1
a	1	-1	-1	-1	2	2	2	2	1
b	-1	1	-1	-1	2	2	2	1	2
ab	1	1	-1	-1	1	1	1	2	2
c	-1	-1	1	-1	2	2	1	2	2
ac	1	-1	1	-1	1	1	2	1	2
bc	-1	1	1	-1	1	1	2	2	1
abc	1	1	1	-1	2	2	1	1	1
d	-1	-1	-1	1	2	1	2	2	2
ad	1	-1	-1	1	1	2	1	1	2
bd	-1	1	-1	1	1	2	1	2	1
abd	1	1	-1	1	2	1	2	1	1
cd	-1	-1	1	1	1	2	2	1	1
acd	1	-1	1	1	2	1	1	2	1
bcd	-1	1	1	1	2	1	1	1	2
abcd	1	1	1	1	1	2	2	2	2

Results and Discussion

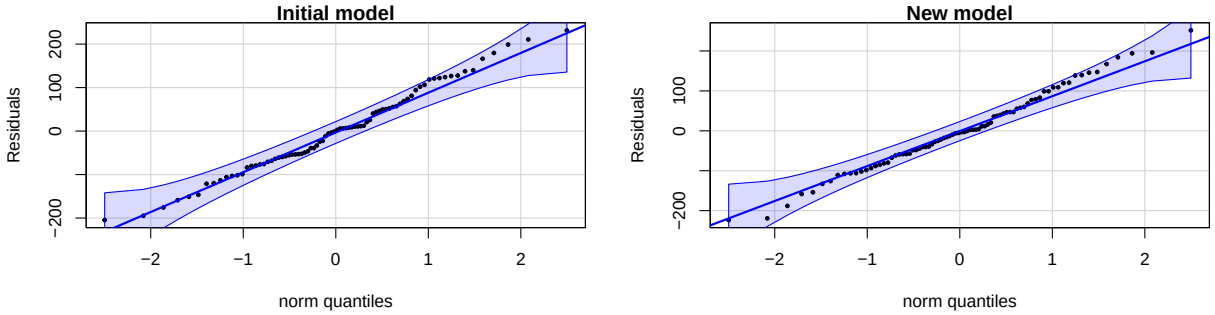
The final factorial design is displayed in Table 1. After inputting and simulating data (*Supplementary Table 1, 4*), we fitted the factors A, B, C, D and the block into a linear regression model. The model summary shows that all the estimated main effects and the more important first-order interactions are positive, which is consistent with our simulation and literature evidence (*Supplementary Table 2*). As seen in Table 2, the total degrees of freedom is $N-1 = 2^4 \times 5 - 1 = 79$. Each main effect and interaction has 1 degree of freedom, summing up to be 15 degrees of freedom for the 15 possible combinations. The remaining degrees of freedom can be partitioned into 4 degrees of freedoms for replicates, 5 degrees of freedoms for blocks within replicates, and 55 degrees of freedoms for error.

Table 2: Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Replicates	49480.029	4	12370.007	-	
Blocks within replicates	28346.734	5	5669.347	-	
A	74602537.307	1	74602537.307	6313.929	<0.001
B	40159610.415	1	40159610.415	3398.878	<0.001
C	29867737.008	1	29867737.008	2527.833	<0.001
D	34937962.848	1	34937962.848	2956.948	<0.001
AB	2524562.299	1	2524562.299	213.664	<0.001
AC	1499792.447	1	1499792.447	126.934	<0.001
AD	99953.551	1	99953.551	8.459	0.005
BC	2229242.429	1	2229242.429	188.67	<0.001
BD	98258.924	1	98258.924	8.316	0.005
CD	212768.431	1	212768.431	18.007	<0.001
ABC (from replicates I, III, IV, and V)	4265.617	1	4265.617	0.361	0.55
ABD (from replicates I, II, IV, and V)	409.895	1	409.895	0.035	0.853
ACD (from replicates I, II,III, and V)	6553.047	1	6553.047	0.555	0.459
BCD (from replicates I, II, III, and IV)	1613.355	1	1613.355	0.137	0.713
ABCD (from replicates II, III, IV, and V)	1542.319	1	1542.319	0.131	0.719
Error	667072.258	55	12128.587		
Total	186991708.911	79			

We then calculated the sums of squares using the expressions specified in the Methods section, and divided them by the corresponding degree of freedoms to obtain mean squares, as seen in Table 2. It appears that

A, B, C, D, AB, AC, AD, BC, BD, and CD have p-values smaller than a significance level of 0.05, so we may conclude that the main effects of A, B, C, D and their first-order interactions are important to total tourist expenditure. This is expected based on our data simulation and the idea that lower-order interactions are of higher importance compared to higher-order interactions. Then, we fitted the significant main effects, their first-order interactions and the block effect into a new model. Using a significance level of 0.05, resulted ANOVA table suggests that they are indeed all significant with p values for all main effects and interactions smaller than 0.05 except for block. However, among all interactions, AB, AC and AD have the smallest p-values ($p < 0.001$), which means that they will still be considered as significant even if we use a smaller significance level (*Supplementary Table 3*). Factor A represents foreign nationality, so this is in line with our simulation and literature evidence, where foreign nationality was considered as an important factor in tourist expenditure (Marrocu, Paci & Zara, 2015; see also Almeida & Garrod, 2017; Vieira & Santos, 2012).



To check the results' validity, we examined the assumptions for both initial and new linear models using residual plots and Normal quantile-quantile plots. Upon inspection, there appears to be no particular pattern such as fanning in the residual plots for both models (*Supplementary Figure 1 & 2*). However, although points in both Normal quantile-quantile plots above are mostly lying within the boundaries around the line, the points in the new model Q-Q plot appear to align closer with the diagonal line than the points from the initial model. Thus, we could conclude that there are no obvious violations of assumptions in both models, but the new model seems to be better because it satisfies the normality assumption better. Therefore, based on the results, we can conclude that foreign nationality, travelling with family, first-time visiting, activity level, and their first-order interactions are significant to the total tourist expenditure.

Conclusion

In conclusion, this study set out to investigate which main effects and interactions of factors foreign nationality, family travelling, first-time visit and activity level have significant impact on the total tourist expenditure, through a 2^4 factorial design with partial blocking and data simulation. The results suggest that all four main effects and their first-order interactions, especially for the interactions that involve foreign nationality, are significant, which are in line with our design, data simulation and literature evidence. Also, we failed to reject our hypothesis that these factors are important to the total tourist expenditure.

References

Almeida, A. & Garrod, B. (2017). Insights from analysing tourist expenditure using quantile regression. *Tourism Economics: the Business and Finance of Tourism and Recreation, 23*(5), 1138–1145. https://doi.org/10.1177/1354816616668108

Marrocu, E., Paci, R., & Zara, A. (2015). Micro-economic determinants of tourist expenditure: A quantile regression approach. *Tourism Management, 50*, 13–30. https://doi.org/10.1016/j.tourman.2015.01.006

Montgomery, D. C. (2020). Design and analysis of Experiments (10th ed.). Wiley.

Santos, & Vieira, J. C. (2012). An Analysis of Visitors' Expenditures in a Tourist Destination: OLS, Quantile Regression and Instrumental Variable Estimators. *Tourism Economics: the Business and Finance of Tourism and Recreation, 18*(3), 555–576. https://doi.org/10.5367/te.2012.0133