Factorial Experiment Design on Key Factors Affecting BMI

Zichen Gong 1005682469

08/04/2022

Introduction

The objective of this experiment is to investigate whether factors potentially leading to heart disease will affect each other. More specifically, **either one of kidney disease, smoking behavior, alcohol-drinking behavior, physical activity and sleep time or an interaction between them collectively are going to affect Body Mass Index (BMI), using an unreplicated 2^5 factorial experiment with 2 blocks and highest order interaction with blocks**. BMI is a good indicator of obesity which is a potential threat over people's health status [1]. Although there is minor difference of BMI standard between men and women, a BMI of 22 is ideal overall. If we manage to prove that some factors will affect BMI, this means that the probability of heart disease for people who have these symptoms/behaviors can be even higher or lower indicated by these effects.

Materials and Methods

Experimental Design and Data

The dataset used for this experiment is composed of a binary variable indicating whether one observation has heart disease or not, and a set of other variables, either numerical or descriptive, indicating potential factors that are highly associated with heart disease [5]. This data originally comes from Centers for Disease Control and Prevention (CDC), collected via telephone surveys from U.S. residents belonging to the Behavioral Risk Factor Surveillance System (BRFSS). Established in year 1984, the most recently modified version we used for this experiment, including data from year 2020, is updated in year 2022, consisting of 401958 rows and 279 columns. Pytlak selected the most relevant factors, and so right not the dataset consists of 319795 observations and 23 columns.

In this design, there are five factors under investigation in total: four binary variables indicating whether one observation has kidney disease or not (factor A), whether the observation smokes or not (factor B), whether the observation drinks alcohol or not (factor C), whether the observation practices physical activities or not (factor D), and one numerical variable indicating sleep time of the observation (factor E).

An unreplicated factorial design experiment using blocking and confounding is implemented. Due to limited resource and time consuming concern, a 2^5 factorial design, which contains 32 treatment combinations is not realistic. If so, this means we need to run 5 main factors, 10 two-factor interactions, 10 three-factor interactions, 5 four-factor interactions and 1 five-factor interaction. It takes time to find people with certain combinations of symptoms/behaviors. Even if using existed observations in the dataset, not everyone is willing to accept follow-up telephone surveys.

Instead, we split each factor into two levels, high level and low level. Observations with high level under one specific factor are marked as 1, while those with low level are marked as -1. For factor kidney disease, observations that have kidney disease are marked as 1, and those who do not have are marked as -1. For factor smoking behavior, observations that smoke are marked as 1, and those who do not smoke are marked as -1. For factor alcohol-drinking behavior, observations that drink alcohol are marked as 1, and those who do not drink are marked as -1. For factor physical activity, observations who do physical activities are marked as 1, and those who do not are marked as -1. For factor sleep time, observations that sleep greater than or equal to 7 hours per day are marked as 1, and those who do not meet this standard are marked as -1. The recommended range of sleep time for adults is within 7 to 9 hours, inclusive [3]. That is why I choose 7h/day as the threshold of dividing high and low levels for the factor sleep time.

When we confound with the highest interaction term, $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1$, the highest order interaction turns out to be $A^{\alpha_1}B^{\alpha_2}C^{\alpha_3}D^{\alpha_4}E^{\alpha_5} = ABCDE$. $L = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5$ gives $L = x_1 + x_2 + x_3 + x_4 + x_5$. x_is are binary. If it is high level, $x_i = 1$. Or if it is low level, $x_i = 0$. For L mod (2) = 0, it is assigned to Block 1. For L mod (2) = 1, it is assigned to Block 2. Instead of running a total of 32 treatment combinations, I can choose either one of the two blocks.

For the factorial design, I need to pick 32 matched observations without partial aliasing. I first create a 2^5 factorial design without randomization using `FrF2()`. Then I pair it up with the original dataset after high level and low level cleaning. Only observations with matched factors observations are kept, and replicated observations are removed. Only 32 observations left at last.

Statistical Analysis

The effect model of this experiment is $\mu_{abcde} = \mu + \tau_a + ... + \tau_e + \tau_{ab} + ... + \tau_{de} + ... + \tau_{abc} + ... + \tau_{cde} + \tau_{abcd} + ... + \tau_{bcde} + \tau_{abcde}$. μ is the general mean. $\tau_a ... \tau_{abcde}$ are main effects and interactions. The null hypothesis is $\tau_a = ... = \tau_e = 0$ for main effects and $\tau_{ab} = ... = \tau_{bcde} = \tau_{abcde} = 0$ for interactions. The alternative hypothesis in this case is at least one of these effects is not equal to 0.

This experiment uses Anova test and Daniel plot to decide whether we should keep some active effects and interactions and which of them are statistically significant.

Table 1: Table 1: Effect Estimates for the Blocked 2ˆ5 Design

	Regression Coefficient	Effect Estimate	Sum of Squares	Percent Contribution
(Intercept)	30.77	61.54	23.81	1.60
A1	0.86	1.72	62.44	4.21
B1	-1.40	-2.79	58.05	3.91
C1	1.35	2.69	52.74	3.55
D1	-1.28	-2.57	112.43	7.57
E1	-1.87	-3.75	32.16	2.17
A1:B1	1.00	2.01	23.32	1.57
A1:C1	0.85	1.71	55.60	3.74
B1:C1	1.32	2.64	2.59	0.17
A1:D1	-0.28	-0.57	71.88	4.84
B1:D1	1.50	3.00	0.95	0.06
C1:D1	0.17	0.35	8.82	0.59
A1:E1	-0.53	-1.05	2.11	0.14
B1:E1	0.26	0.51	66.99	4.51
C1:E1	-1.45	-2.89	1.46	0.10
D1:E1	-0.21	-0.43	149.82	10.09
A1:B1:C1	-2.16	-4.33	269.00	18.12
A1:B1:D1	2.90	5.80	21.16	1.42
A1:C1:D1	0.81	1.63	43.80	2.95
B1:C1:D1	1.17	2.34	11.42	0.77
A1:B1:E1	-0.60	-1.19	49.90	3.36
A1:C1:E1	-1.25	-2.50	36.68	2.47
B1:C1:E1	-1.07	-2.14	2.73	0.18

	Regression Coefficient	Effect Estimate	Sum of Squares	Percent Contribution
A1:D1:E1	0.29	0.58	1.60	0.11
B1:D1:E1	0.22	0.45	0.82	0.06
C1:D1:E1	-0.16	-0.32	126.64	8.53
A1:B1:C1:D1	1.99	3.98	69.03	4.65
A1:B1:C1:E1	-1.47	-2.94	0.48	0.03
A1:B1:D1:E1	-0.12	-0.24	10.19	0.69
A1:C1:D1:E1	0.56	1.13	29.95	2.02
B1:C1:D1:E1	-0.97	-1.93	86.26	5.81
A1:B1:C1:D1:E1	-1.64	-3.28	0.00	0.00

The coefficient estimates generated via Anova, the effect estimates, sum of squares and the percent contribution of each sum of square is demonstrated on Table 1. The effect estimates of main effects, 2-factor to 4-factor interactions should be identical to the effect estimates without block effects. If we briefly take a look at coefficient estimates, main effect E, interaction effect between ABC, interaction effect between ABD, and interaction effect between ABCD all have larger absolute values, indicating that they might be important.

In this experiment, the interaction estimate of highest order interaction ABCDE is the sum of the original interaction effect and the block. It is also equal to the block effect, which can be calculated through the difference between means of response variables for 2 blocks. The block effect is $B\bar{M}I_{\text{block 1}} - B\bar{M}I_{\text{block 2}} = 32.41 - 29.13 = 3.28$. In other words, block effect is equal to the sum of interaction effect and the block.

Results and Discussion

Table 2: Table 2: Analysis of Variance

	Degree of Freedom	Sum of Squares	Mean Square	F Value	P Value
E	1	112.425012	112.425012	4.4559795	0.0519755
A	1	23.805000	23.805000	0.9435142	0.3467795
B	1	62.440312	62.440312	2.4748296	0.1365338
C	1	58.050312	58.050312	2.3008314	0.1500949
D	1	52.736450	52.736450	2.0902158	0.1688145
A:B	1	32.160200	32.160200	1.2746736	0.2766250
A:C	1	23.324450	23.324450	0.9244675	0.3515490
B:C	1	55.598512	55.598512	2.2036540	0.1583934
A:D	1	2.587813	2.587813	0.1025683	0.7531870
B:D	1	71.880050	71.880050	2.8489748	0.1121087
C:D	1	0.952200	0.952200	0.0377406	0.8485715
A:B:C	1	149.818050	149.818050	5.9380572	0.0277506
A:B:D	1	269.004012	269.004012	10.6620078	0.0052174
A:C:D	1	21.157512	21.157512	0.8385807	0.3742907
B:C:D	1	43.804800	43.804800	1.7362088	0.2073914
A:B:C:D	1	126.643613	126.643613	5.0195355	0.0406275
Residuals	15	378.452187	25.230146	NA	NA

After consulting via Daniel plot, I pick main effect E, interaction effect between ABC, interaction effect between ABD, and interaction effect between ABCD as the final effect terms. This fits what we observed from Table 1. The normality assumption and constant variance assumption of residuals are verified.

If we take a significance level of $\alpha = 0.1$, based on Table 2, the final Anova output, main effect E, interaction effect between ABC, interaction effect between ABD, interaction effect between ABCD all have enough evidence to reject the null hypothesis. This indicates that, sleep time itself, interaction between kidney disease, smoking and alcohol drinking behavior, interaction between kidney disease, smoking and physical activity, and interaction between kidney disease, smoking, drinking alcohol and physical activity, all have strong effects on BMI.

Conclusion

Using an unreplicated 2^5 factorial design experiment with 2 blocks and the highest order interaction confounded, we manage to find out that the main factor sleep time, the interaction effect between whether people have kidney disease, smoke and drink alcohol, the interaction effect between whether people have kidney disease, smoke and do physical activity, and the interaction effect between whether people have kidney disease, smoke, drink alcohol and do physical activity all have strong enough evidence to affect people's BMI. We can design further studies investigating how solid these effects are. For a model predicting heart disease, if the model consists of these factors and BMI at the same time, with careful calculation, we might consider removing BMI from the model for simplicity.

References

1. Are BMI charts different for men & women? (2019). Retrieved April 2, 2022, from https://gasparinutrition.com/blogs/fitness-facts/are-bmi-charts-different-for-men-women#:~:text=Recent%20studies%20have%20found%20that,21%20for%20women%20is%20healthy.

2. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr (Last Accessed: April 3, 2022)

3. How Much Sleep Do I Need? (2017). Retrieved April 2, 2022, from https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html

4. John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/ (Last Accessed: April 3, 2022)

5. Kaggle. (2022). *Personal Key Indicators of Heart Disease* [Data file]. Retrieved April 2, 2022, from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

6. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (Last Accessed: April 3, 2022)

7. Ulrike Gr"omping (2014). R Package FrF2 for Creating and Analyzing Fractional Factorial 2-Level Designs. Journal of Statistical Software, 56(1), 1-56. URL http://www.jstatsoft.org/v56/i01/ (Last Accessed: April 3, 2022)

8. Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.37. (Last Accessed: April 3, 2022)

The report is constructed using `R-4.1.2`.