This data originally was collected via telephone surveys from U.S. residents belonging to the Behavioral Risk Factor Surveillance System (BRFSS). The data was established in year 1984. The most recently modified version including data from year 2020, is updated in year 2022.

The objective 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. 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.

Due to limited resource and time consuming concern, a 2^5 factorial design, which contains 32 treatment combinations is not realistic. 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.

When we confound with the highest interaction term, the highest order interaction turns out to be ABCDE. We modulate L by 2 and put observations into block 1 and block 2.

For the factorial design, I pick 32 matched observations without partial aliasing. I first create a 2 to the power of 5 factorial design without randomization. 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. 32 observations left at last.

The coefficient estimates generated via Anova, the effect estimates, sum of squares and the percent contribution of each sum of square is demonstrated in 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.

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.

If we take a significance level of α = 0.1, based on Table 2, the final Anova output, 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.

The normality assumption and constant variance assumption of residuals are verified.

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.