individ_formal_analysis

Trevor Kwan

20/04/2021

Missing Data

Of the 1355 observations, 317 were missing so the first step of the formal analysis was to determine how to handle the missing data. A complete cases method, where all missing observations were removed from the dataset was first attempted, but this required a strong assumption for the data to be missing completely at random. Because there was no information provided on how the scores were missing, this assumption made the complete cases method undesirable. A multiple imputation method using predictive means matching to estimate missing values was then conducted. This was done through a multiple imputation by chained equations (MICE) implementation in R [@RCore2021; @mice]. The resulting dataset from multiple imputation was used for the formal analysis.

Means Comparison

Preliminary exploratory analyses suggests GSI scores of the treatment group to be higher than GSI scores of the control group. Using the full imputed dataset, a two-sample t-test found this difference to be non-significant. Examination of normality assumptions using the Shapiro-Wilk test found both the distribution of the treatment group and the distribution of the control group to violate the normality assumption. Thus, a two-sample Wilcoxon test was required, and the test found the GSI difference in treatment groups to be significant.

Table 1: Results of t-tests and wilcoxon tests for treatment vs. control group comparison

Estimated Difference	Standard Error	p-value (t-test)	p-value (Wilcoxon test)
-0.053	0.037	0.153	0.009

Mixed Effects Models

These mean comparison tests found non-significant differences between the two treatment groups, but a mixed-effects models approach was taken to account for the large variations between individuals and appropriately model the clustered nature of the data, thus allowing for individual specific inference as well as population average inference. This allowed the correlation between within-subject measurements to be modeled appropriately. Mixed-effects models also automatically incorporate missing data, meaning the longitudinal data need not be balanced and each individual may have different repeated measurements. To investigate the difference in GSI scores between treatment groups, a mixed-effects model was constructed, taking the form:

$$GSI_{ijkl} = \beta_0 + \beta_1 treatment_{ijkl} + subject_i + month_j + gender_k + education_l + \epsilon_{ijkl} \quad (1)$$

where GSI is the mental distress score, *treatment* is an indicator of the treatment group, *subject* is the random effect for the i^{th} subject, *month* is the random effect for the j^{th} month, *gender* is the random effect for the k^{th} gender, *education* is the random effect for the l^{th} education level, and ϵ_{ijkl} is the random error. This model showed non-significant decreases in GSI when switching from the treatment group to the control group (Table 4, see appendix for full summary).

Table 2: Estimated mean scores and p-values associated with parameters (intercept, and parameter estimates for treatment vs. control groups) for the fitted mixed model of the form shown in Equation 1

Group	Estimated Mean Score	Parameter Estimate	Standard Error	p-value
Treatment	0.918	0.918	0.208	0.011
Control	0.877	-0.041	0.059	0.657

Mixed-Effects Regression

The previous model only provided information on whether the treatment groups were significantly different in GSI scores, but did not specifically investigate whether or not subjects' mental distress in both of the treatment groups decreased significantly over time. To examine the effect of month on GSI scores, a mixed-effects regression approach was also taken, taking the form:

$$GSI_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})month_{ij} + b_{2i}gender_{ij} + \epsilon_{ij} \tag{2}$$

but this time the month variable is the fixed effect and the variance between month and gender were shared and varied by subject. This model showed that there was a significant decrease in GSI score of approximately 0.005 for each additional 1 month increase (Table 5, see appendix for full summary). Examining the residuals from the regression showed slight evidence of heteroscedasticity, but little evidence of violations of the normality assumption for this model (see appendix).

Table 3: Parameter estimates and p-values for the fitted mixed-effects regression model of the form shown in Equation 2

Parameter	Estimated Value	Standard Error	p-value
Mean GSI at Month = 0	0.993	0.034	< 0.001
Estimated Per- Month Change in GSI	-0.005	0.001	< 0.001