EPSY 905 Multivariate Statistics Homework:

Project Assignment #1

Aaron Matthew Simmons and Dasha Yermol

Department of Psychology, University of Kansas

EPSY 905

Dr. John Poggio

March 18, 2024

**QUESTION AREA 1:**

**Background and Methods**

A 2 x 3 between-subjects multivariate analysis of covariance (MANCOVA) was performed to assess how the type of activity (skiing versus snowboarding) and the lesson type (private, group, no lesson) affected the Time and the number of falls it took to travel down a hill. Adjustments were made for the one covariate: the Socio-Economic Status (SES) of each participant, which was coded on a 5-point interval scale from 1 (very low income) to 3 (median) to 5 (very high income).

The *R* statistical language was used for the analyses. In this sample, there are no missing data, and the total sample size was 594 before trimming. However, there are unequal samples across groups: 112 for Ski/Private Lesson (Group 1), 121 for Ski/Group Lesson (Group 2), 64 Ski/No Lesson (Group 3), 112 for Snowboard/Private Lesson (Group 4), 121 for Snowboard/Group Lesson (Group 5), and 64 for Ski/No lesson (Group 6).

***Outliers:***

We assessed outliers after separating each data among the six groups. The raw data consisted of one univariate outlier was found using a criterion z = , ⍺ = 0.001, which was Case ID #587 (Group 5: Snowboard/No Lesson) in terms of Time. This same data point Case ID #587 was also a multivariate Mahalanobis D2 outlier using a criterion of ⍺ = 0.001 and df = 2 with a critical . Simply deleting this case caused continued outliers within this group.

To account for this outlier, we attempted winsorizing time values across all groups such that extreme time values less than 1% or greater than 99% of the data were replaced by their lowest and highest untrimmed time values, respectively. However, this winsorizing approach affected the assumption of homogeneity of variance-covariance via Box’s M test (, and follow-up Levene’s test found evidence to reject the null hypotheses (i.e., p < 0.001) which suggests a lack of homogeneity of variance in both dependent variables. So instead of winsorizing, we conducted iterative case-wise deletion of outliers in this group resulting in deletion of Case ID #587, #7, #525, and #283 until univariate and multivariate outlier analysis became satisfactory. After case-wise deletions, the sample size of Group 5 was reduced 60.

Therefore, the total sample size (N = 594) of the raw data was reduced to 590 with the deletion of outliers within one of the six groups. After being trimmed, there were no univariate or multivariate within-cell outliers at ⍺ = 0.001

**Multivariate Normality**

The sample size of 590 includes over 60 data points for each cell of a 2 x 3 between-subjects design which is more than the 20 degrees of freedom for error suggested to assume multivariate normality of the sampling distribution of means, even with unequal sample sizes; there are far more cases than dependent variables in the smallest cell.

***Multicollinearity***

There is a correlation between Time and Falls of r = 0.69; given this, MANCOVA is appropriate since it is not greater than 0.9. It is also sensible that the direction of this association is positive since the amount of time it takes to travel down a hill would be affected by how many falls took place (e.g., the more falls, the longer it took to get down the hill).

***Linearity***

The linearity assumption holds for this data and was assessed via scatterplots between Time and Falls for each group (see Figure 1).

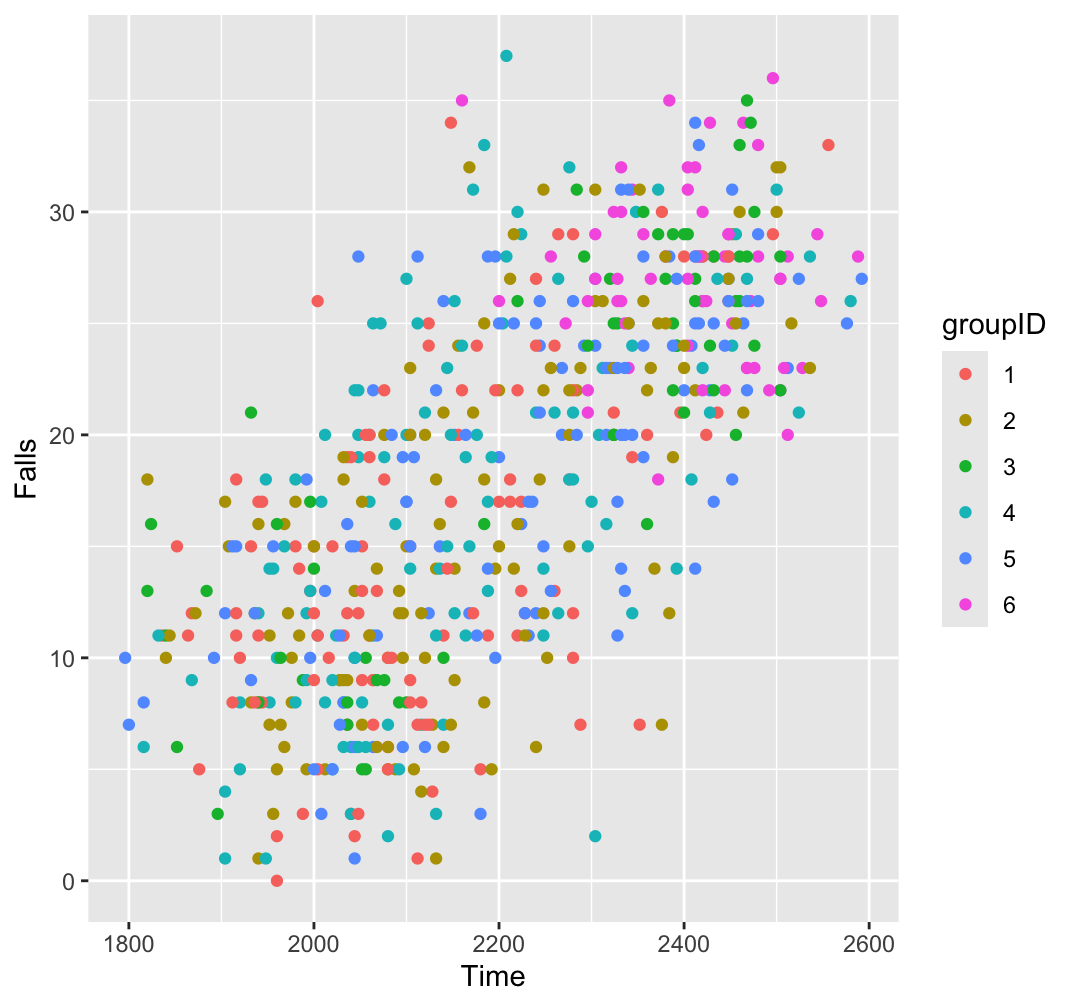


Figure 1: Scatterplots for each group to assess linear form.

***Homogeneity of Variance-Covariance***

Box’s M test was performed on the trimmed data and showed a lack of homogeneity of variance-covariance ( Follow-up Levene’s tests similarly found evidence to reject the null hypotheses (i.e., p < 0.001) which suggests a lack of homogeneity of variance in both dependent variables. To adjust, we attempted to run square root and logarithmic transformations on one or both of the dependent variables, but found similar heterogeneity of variance-covariance. Instead, to keep the interpretation of analyses more tractable, we decided to not transform the data and focus subsequent MANCOVA analysis using Pillai’s trace since it is most robust to assumption of violations; in this case, due to unequal sample sizes and heterogeneity of variance-covariance.

In summary, when evaluating MANCOVA assumptions, the assumptions of normality, normality, multicollinearity, and linearity was satisfactory. However, there was not a satisfactory assumption of the homogeneity of variance-covariance matrices and attempts to transform the data did not aid in these assumptions. Therefore, we did not transform the data and will use Pillai’s trace as a multivariate statistic since it is most robust to assumption of violations; in this case, due to unequal sample sizes and heterogeneity of variance-covariance.

**Results (RQ answers for HW#1 are bolded):**

With the use of Pillai’s trace, the combined dependent variables of time and number of falls were significantly related to activity, *F*(2, 582) = 8.21, *p* < 0.001 and to lesson type, *F*(4,1166) = 15.11, *p* < 0.001, and to the SES covariate, *F*(2, 582) = 186.44, *p* < 0.001, but not the interaction of activity and lesson, *F*(4, 1166) = 1.95, p = 0.10 (**RQ#1 & RQ#2)**. There were small associations between the dependent variables and the main effects of type of activity and lesson type, η2 = 0.03, 95% CI = [0.01, 1.00] and η2 = 0.05, 95% CI = [0.03, 1.00], respectively. However, there was a much larger association between the dependent variables and the SES covariate, η2 = 0.39, 95% CI = [0.34, 1.00].

Since the MANCOVA main effects were significant, it is appropriate to consider univariate and post-hoc procedures to better understand the relationships embedded in the multivariate model. For example, the univariate effects of each independent variable on the dependent variables, after controlling for SES, showed significant effects (p < 0.001); i.e., when each independent variable was regressed on both dependent variables separately, after controlling for SES, showed significant between-subject effects. However, all effect size values were small, and no interaction effects were significant (see Table 1 for SPSS table) (**RQ#4**).

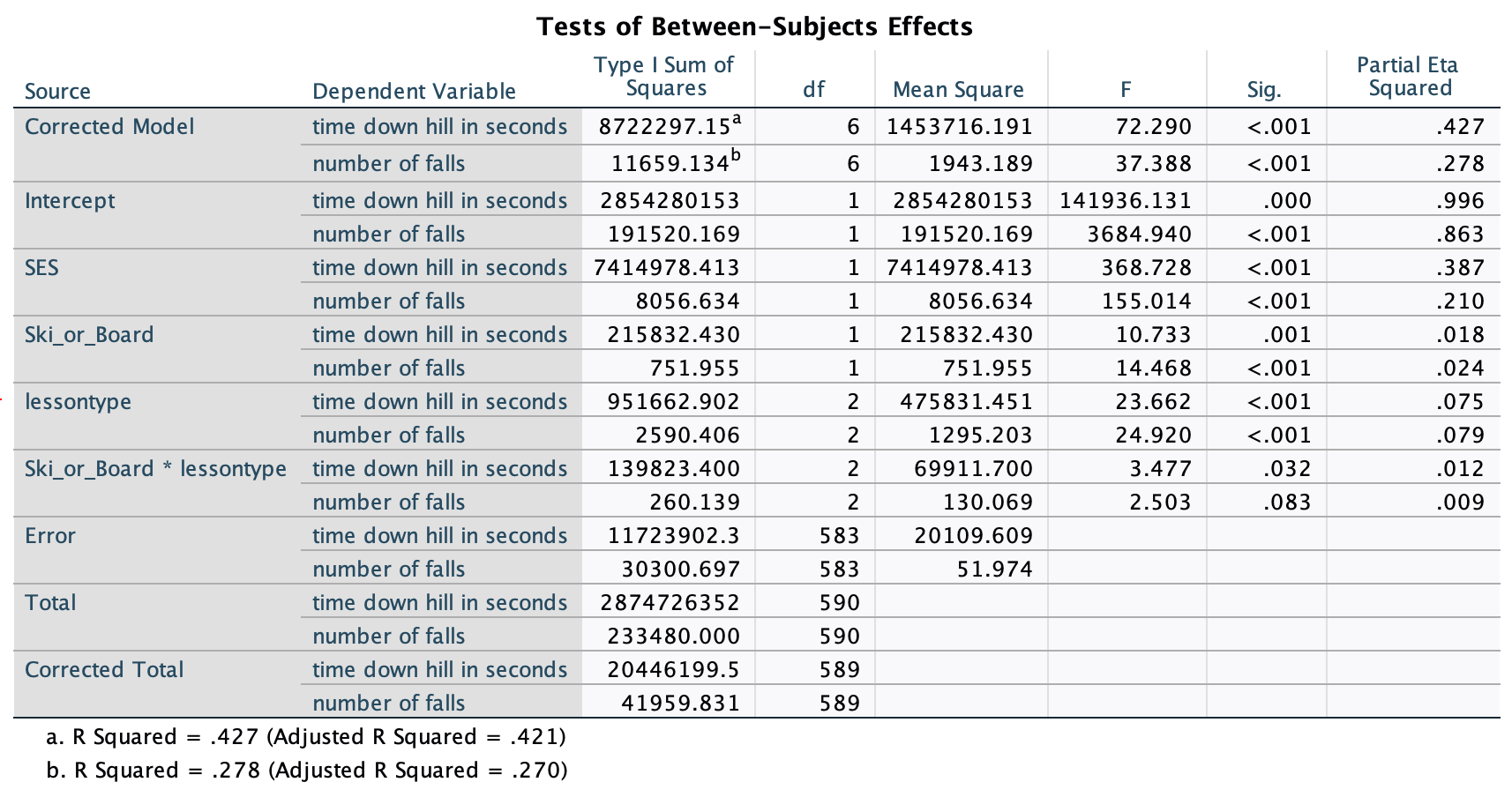


Table 1: Univariate Analysis of Covariance from SPSS (*R* was limited in providing this table)

As a post-hoc procedure to determine which training produced the best results for skiing and snowboarding, the time it took to travel down the hill was less for skiers (*M* = 2185.68, *SE* = 8.56, 95% CI = [2168.84, 2202.52]) than snowboarders (*M* = 2237.50, *SE* = 8.81, 95% CI = [2220.20, 2254.81]), and the number of falls was less for skiers (*M =* 17.24, *SE* = 0.44, 95% CI = [16.38, 18.10]) than snowboarders (*M* = 20.12, *SE* = 0.45, 95% CI = [19.23, 21.00]). The results of this specific analysis in shown in Figure 2 (**RQ#3 & RQ#4**).

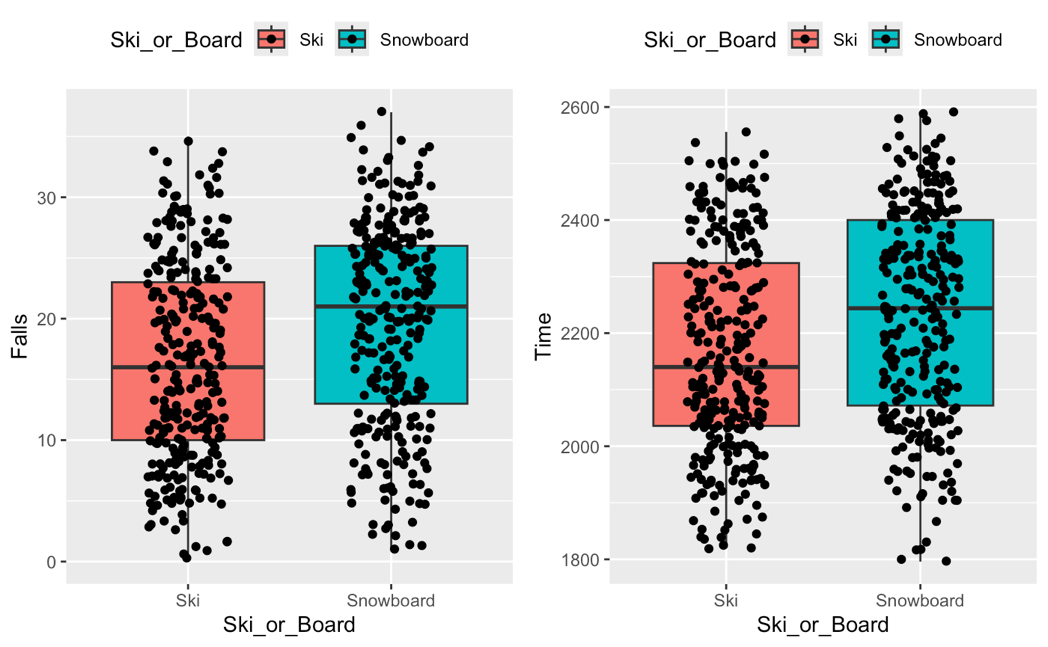


Figure 2: Ski vs. Snowboard results on Falls and Time down a hill.

**Discussion**

We have shown with MANCOVA that there is evidence that the type of activity (snowboarding or skiing) and the type of lessons one received (private, group, or none) is significantly related to the combined dependent variables of the number of falls and time it took to travel down a hill (but not their interaction). Further post-hoc procedures showed that snowboarding is associated with more falls and time down the bottom of the hill than skiing.