EPSY 905 Multivariate Statistics Homework:

Project Assignment #1

Aaron Matthew Simmons and Dasha A. Yermol

Department of Psychology, University of Kansas

EPSY 905

Dr. John Poggio

March 18, 2024

<u>QUESTION AREA 1:</u>

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 = |3.3|, α = 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 D^2 outlier using a criterion of α = 0.001 and df = 2 with a critical $\chi^2 = 13.82$. 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 ($\chi^2 =$

$94.087, \mathrm{df} = 15, \mathrm{p} < 0.001$), 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 Group 5 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 for multivariate analyses as a result of the deletion of outliers within one of the six groups. After being trimmed, there were no univariate or multivariate within-cell outliers at $\alpha = 0.001$ across any groups.

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 r = 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).

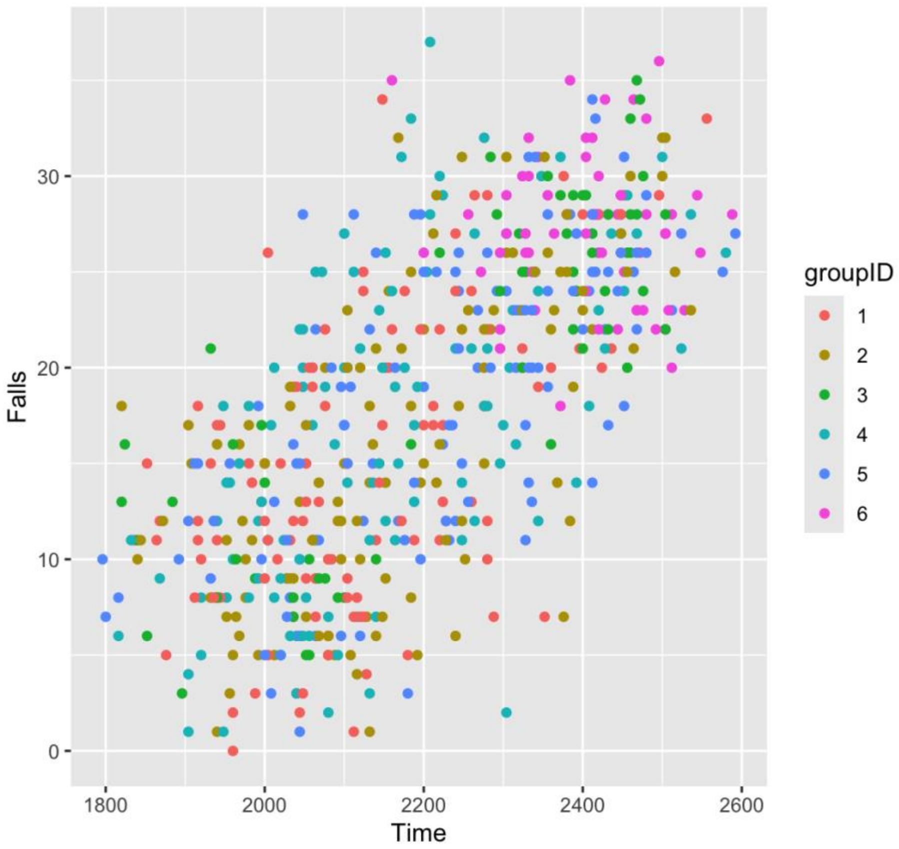


Figure 1: Scatterplots for each group to assess linear form.

Homogeneity of Variance-Covariance

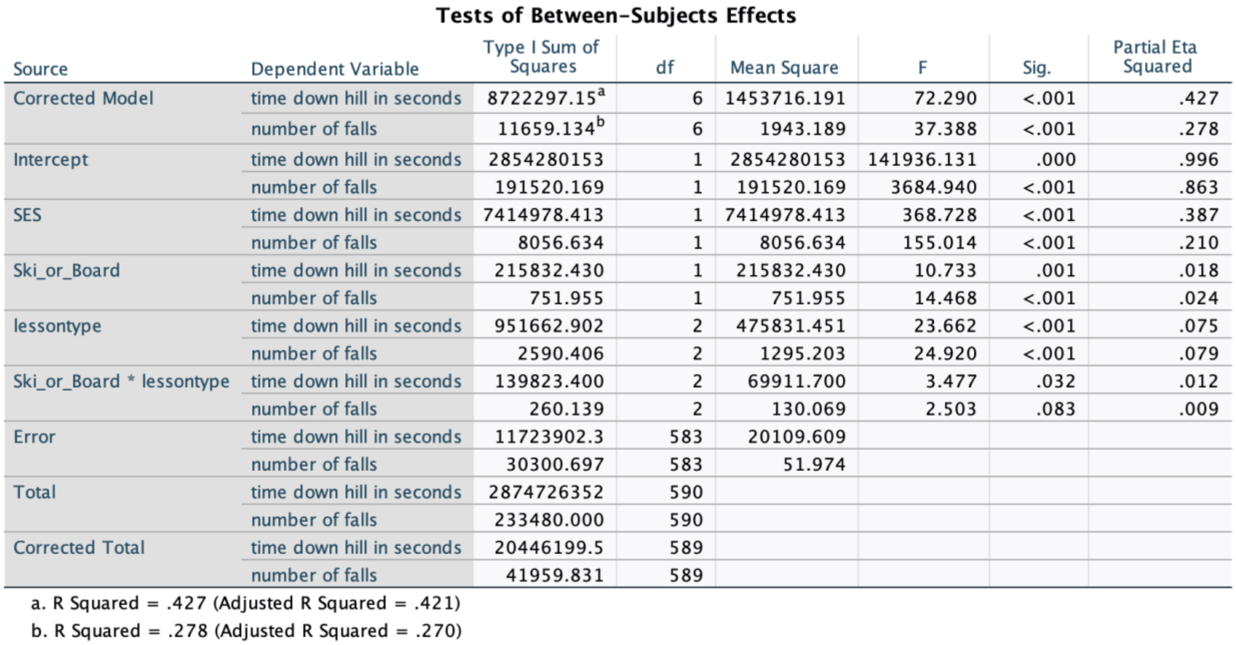
Box's M test was performed on the trimmed data and showed a lack of homogeneity of variance-covariance ($\chi^2 = 85.906, \text{df} = 15, \text{p} < 0.001$). Follow-up Levene's tests similarly found evidence to reject the null hypotheses (i.e., p < 0.001) which suggests a lack of homogeneity of variance 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, $F(4, 1166) = 1.95$, $\mathrm{p} = 0.10$ (**RQ#1 & RQ#2)**. There were small associations between the dependent variables and the main effects of type of activity and lesson type, $\eta^2 = 0.03$, 95% CI = [0.01, 1.00] and $\eta^2 = 0.05$, 95% CI = [0.03, 1.00], respectively. However, there was a much larger association between the dependent variables and the SES covariate, $\eta^2 = 0.39$, 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 ($\mathrm{p} < 0.001$); 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**).



Tests of Between-Subjects Effects

Source	Dependent Variable	Type I Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	time down hill in seconds	8722297.15[a]	6	1453716.191	72.290	<.001	.427
	number of falls	11659.134[b]	6	1943.189	37.388	<.001	.278
Intercept	time down hill in seconds	2854280153	1	2854280153	141936.131	.000	.996
	number of falls	191520.169	1	191520.169	3684.940	<.001	.863
SES	time down hill in seconds	7414978.413	1	7414978.413	368.728	<.001	.387
	number of falls	8056.634	1	8056.634	155.014	<.001	.210
Ski_or_Board	time down hill in seconds	215832.430	1	215832.430	10.733	.001	.018
	number of falls	751.955	1	751.955	14.468	<.001	.024
lessontype	time down hill in seconds	951662.902	2	475831.451	23.662	<.001	.075
	number of falls	2590.406	2	1295.203	24.920	<.001	.079
Ski_or_Board * lessontype	time down hill in seconds	139823.400	2	69911.700	3.477	.032	.012
	number of falls	260.139	2	130.069	2.503	.083	.009
Error	time down hill in seconds	11723902.3	583	20109.609			
	number of falls	30300.697	583	51.974			
Total	time down hill in seconds	2874726352	590				
	number of falls	233480.000	590				
Corrected Total	time down hill in seconds	20446199.5	589				
	number of falls	41959.831	589				

a. R Squared = .427 (Adjusted R Squared = .421)

b. R Squared = .278 (Adjusted R Squared = .270)

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**).

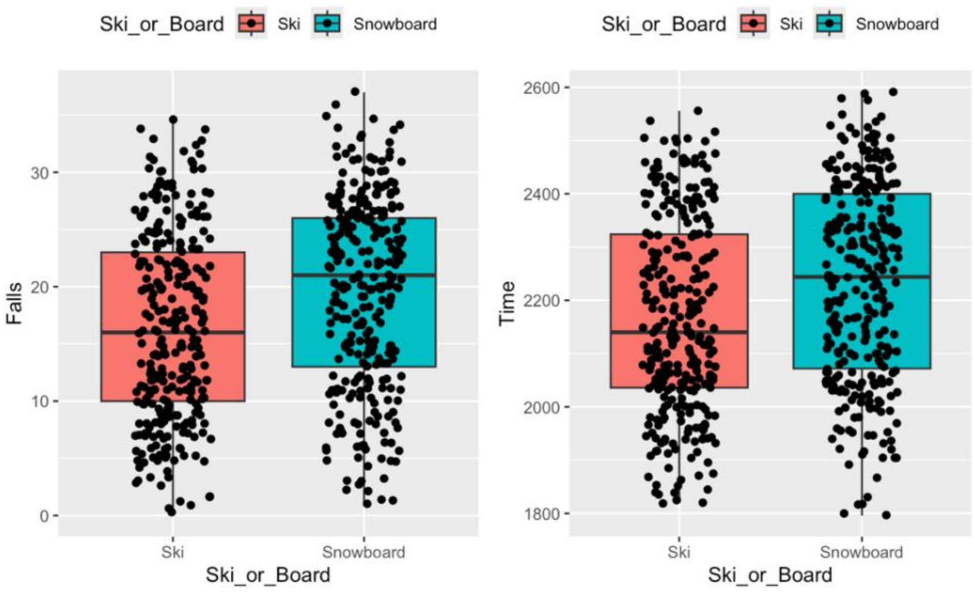


Figure 2: Ski vs. Snowboard results on Falls and Time down a hill.

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.

QUESTION AREA 2:

Background and Methods

A multivariate analysis of variance (MANOVA) was performed to assess how the Gender of the Driver (female vs. male) and the Car Color (red, blue, green, black, white) affected the Response Time and Participant's Response (Blink or Honk). However, a MANOVA should be interpreted with caution in this case because the data violates assumptions of continuous dependent variables (since participant's response of blinking or honking is a binary variable).

The *R* statistical language was used for the analyses. In this sample, there are no missing data, and the total sample size was 422 observations. There are unequal samples across gender of driver groups: 178 females (Group 1) and 244 males (Group 2). However, there were approximately equal samples of car color groups: 84 Red, 83 Blue, 84 Green, 89 Black, 82 White.

Outliers

We assessed univariate and multivariate outliers after separating the dataset among the five car color groups. We used the box plot method to check for univariate outliers. One univariate outlier was found (11.56 Response Time in seconds), however, it was labeled as "not extreme" by the test. Therefore, we decided to keep this outlier in the dataset. Furthermore, we conducted the Mahalanobis Distance, D^2, test to check for multivariate outliers. This test revealed no multivariate outliers with a criterion of $\alpha = 0.001$ and df = 2.

Distribution of Response Time Variable

MANOVA assumes that the dependent variables are normally distributed within each group. As shown in Figure 3, the Response Time variable appears normally distributed both in the Car Color groups (left image) and in the Gender of the Driver groups (right image). Or other dependent variable, whether the participant blinked or honked (“Blink_or_Honk” in dataframe) is a binary variable, so it cannot be normally distributed.

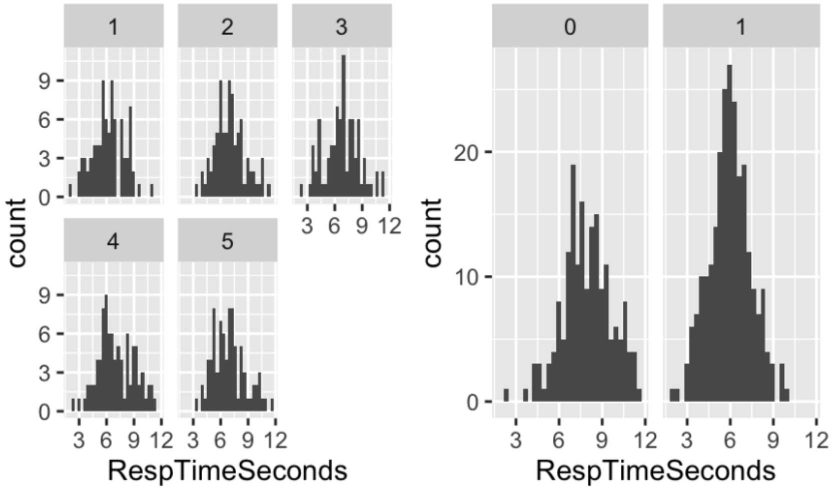


Figure 3: Distribution of Response Time variable by groups: Car Color (left) and Gender of Driver (right)

Univariate Normality

We conducted the Shapiro-Wilks test of univariate normality for the Response Time dependent variable grouped by Car Color (IV #1) and Gender of Driver (IV #2). Results indicate univariate normality for the Response Time variable. When grouped by Car Color, p values ranged from 0.07 to 0.84. When grouped by Gender of Driver, p values were 0.23 and 0.77. According to the Shapiro-Wilks Test, Response Time is normally distributed for each group since the p values were greater than 0.05 for each group. We cannot run the Shapiro-Wilks test on the other DV (Driver's Reaction of Blinking or Honking) because it is a binary/categorical variable. Figure 4 visualizes the univariate normality of the Response Time variable.

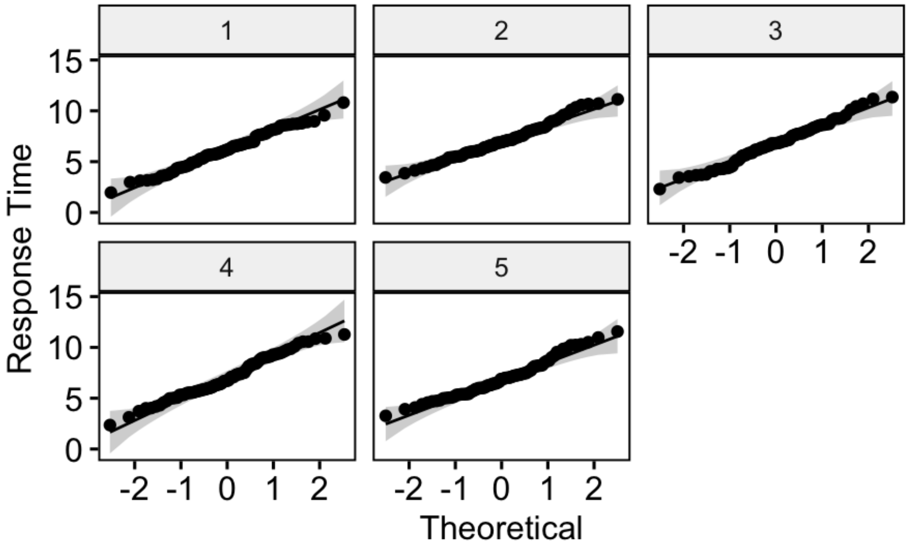


Figure 4: QQ plots show univariate normality of Response Time (DV #1) by Car Color

Multivariate Normality

We conducted the Shapiro test to assess multivariate normality for the Response Time variable (DV #1). The test revealed multivariate normality for the Response Time variable, *p* = 0.051. The p value of the Shapiro test is slightly greater than 0.05, which indicates that it is not significant, and we can assume multivariate normality. We cannot perform this test on the other dependent variable (Driver's reaction of blink or honk) because it is a binary/categorical variable.

Multicollinearity

We have two outcome variables: Response Time and Participant’s Response (Blink or Honk). Since Driver's Reaction is a binary/categorical variable and Response Time is a continuous variable, we will need to conduct a point-biserial correlation to evaluate multicollinearity. The biserial correlation indicates a slight positive correlation between Response Time (DV#1) and Blink or Honk (DV#2), r = 0.035. Since the correlation is not greater than 0.9, we do not have multicollinearity between these two variables.

Linearity

The linearity assumption holds for this data and was assessed via logistic regression and by plotting the relationship between Response Time and the log odds of the outcome, the participant's response (blink or honk), see Figure 5.

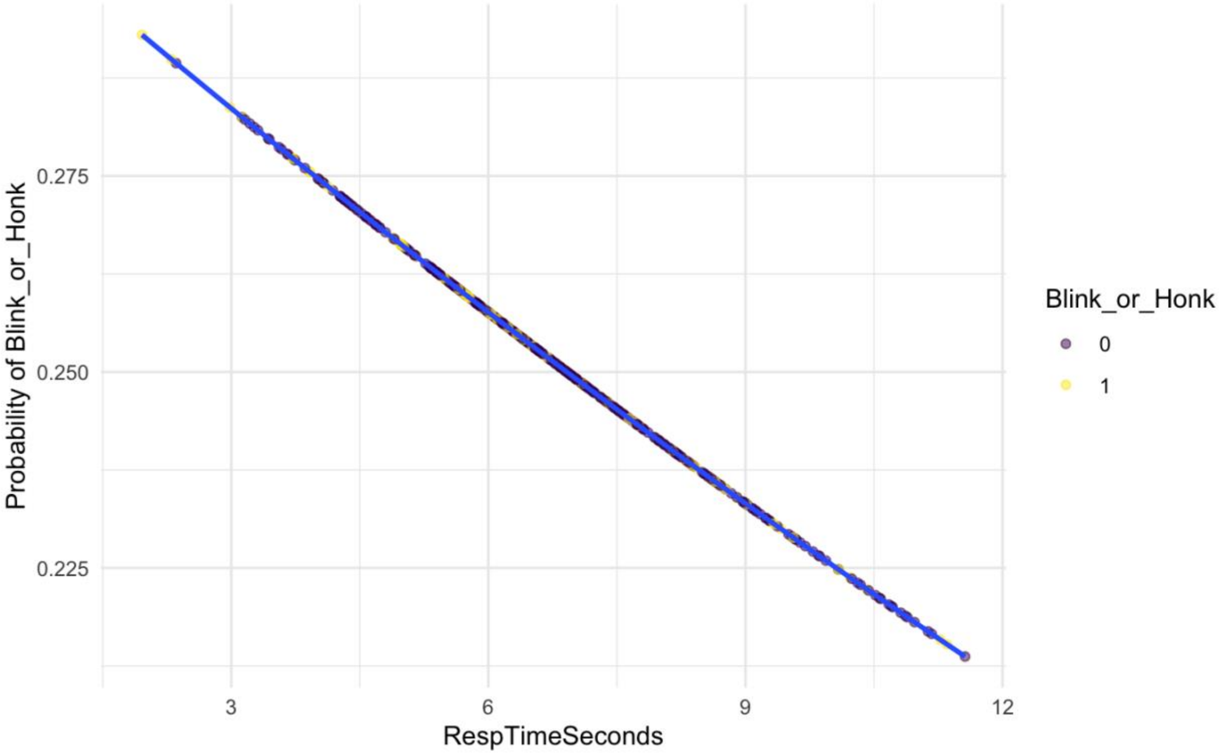


Figure 5: Log odds of Participant's Response by Response Time

Homogeneity of Variances and Covariances

The Levene's test of equality of variances assumes continuous dependent variables, so we are not able to assess the homogeneity of variances in this case because we have one continuous dependent variable (Response Time) and one binary/categorical dependent variable (Blink or Honk response). We are also not able to conduct Box's M test to evaluate homogeneity of covariances because it is designed for continuous variables and depends on the variance of those continuous variables.

Summary of Assumption Checking

In summary, when evaluating MANCOVA assumptions, the assumptions of outliers, univariate normality, multivariate normality, multicollinearity, and linearity were satisfied. However, this data does not satisfy the assumption of homogeneity of variances or covariances. Since one of the dependent variables is binary/categorical, MANOVA is not appropriate to conduct, so we would prefer to use an alternative statistical model, like Structural Equation Modeling, instead. However, we did run an exploratory MANOVA, but we must interpret the results with great caution. We will use Pillai's trace as a multivariate statistic since it is most robust to assumption of violations; in this case, due to the binary/categorical dependent variable.

Results

With the use of Pillai's trace, the combined dependent variables of Response Time and Response of Driver (Blink or Honk) were significantly related to Car Color, F (4, 832) = 4.069, $p < 0.001$ and to Gender of Driver, F (1, 415) = 87.571, $p < 0.001$. Both Car Color and Gender of Driver have significant effects on Response Time and Response of the Driver. However, again, these results should be interpreted with extreme caution because MANOVA is not an appropriate statistical technique when one of the two dependent variables is binary/categorical because MANOVA makes an assumption of continuous dependent variables. We also fit a second MANOVA model with an interaction between the two independent variables. The combined effect of the two independent variables, Car Color and Gender of Driver, did not significantly influence the two dependent variables, Response Time and Participant Response, F (4, 824) = 0.453, $p = 0.89$. These results suggest that the combined effect of the two independent variables may not influence the Response Time and Participant Response Type (Blink or Honk); rather, independently, Car Color and Gender of Driver impact the dependent variables.

Exploratory Structural Equation Modeling

We believe that a Structural Equation Model (SEM) would be more appropriate with this dataset because SEM can handle a binary/categorical dependent variable, unlike MANOVA, which would treat a binary dependent variable as continuous/makes an assumption of continuous dependent variables. We attempted to fit an SEM model with the two independent and two dependent variables, however, the model results had a number of missing values. Since we have not yet learned how to properly fit an SEM model, it was difficult for us to troubleshoot these missing values and we were therefore not able to interpret the results. We look forward to learning more about SEM later on in this course.

Discussion

We have shown with MANOVA that there is evidence that Gender of Driver and Car Color independently influence Response Time and Participant's Response (Blink or Honk). However, when we fit another MANOVA with an interaction between the independent variables, we did not find a significant effect of the Gender of Driver and Car Color on the dependent variables. Although these results should be taken with caution, since MANOVA assumes continuous dependent variables and our data includes a binary dependent variable, they are still suggestive of two independent relationships between each independent variable with the two dependent variables. These results suggest that participants pay attention to the color of the car or the gender of the driver and make an assumption that influences their behavior (via Response Time and Blink/Honk). As a future goal, we hope to explore this data again in an SEM framework.