Name: Md Allama Ikbal Sijan

**Hierarchical Linear Modeling**

Psyc 741, Spring 2025

**Due Date**: February 17, 2025 (11:00 AM)

***Homework #1***

You will be completing the following questions, some of which involve conducting analyses in R/RStudio. You will submit a completed version of this assignment document (i.e., enter your answers into this Word document when needed). Some of the answers involve you copying and pasting your R code. However, you will also submit your *complete* R code (saved as a .R script file).

***Homework #1, Part A: Multiple Linear Regression Review***

For Part A, you will be using data from the **satisfaction.sav** SPSS dataset on Canvas. These data derive from a customer satisfaction survey that was sent out to customers who visited a big-box retail store. Demographics were collected from the customers, information about the potential purchase (e.g., the department in the store, whether the purchase was made or not), as well as satisfaction with various aspects of the shopping experience (price satisfaction, variety satisfaction, organization satisfaction, service satisfaction, item quality satisfaction, and overall shopping satisfaction). A scale satisfaction variable was then created by taking the mean of all these satisfaction items (satisfaction\_mean).

1. Import the satisfaction data set into RStudio.
2. The regression model below contains some categorical predictors: whether the shopper interacted with an employee during the shopping experience (contact) and the primary department the shopping experience was in (dept). So that you will be able to properly interpret the regression coefficients, extract the value labels for these two variables. Even if you have SPSS and can manually open the file to see these value labels, use code in R to find them for this question. [2 points]

**Contact**

No Yes

0 1

**Department**

Clothing Automotive Electronics

1 2 3

1. Run a multiple regression predicting the satisfaction\_mean variable using the lm() function. Enter the shopper’s **age**, the **contact** variable, and the **department** variable as predictors in the model. Copy/paste your model syntax below. [10 points]

model <- lm(satisfaction\_mean ~ age + contact + as.factor(dept), data = df\_s)

summary(model)

1. Report the overall model statistics (e.g., R2, F statistic) below in APA format. [4 points]

The multiple linear regression model predicting the satisfaction level was not statistically significant, *F* (4, 577) = 1.99, *p* = 0.09. The set of independent variables (age, contact, and department) explained approximately 1.36% of the variance in the mean satisfaction score (*R²* = .01362).

1. Regardless of your conclusion in the above question, interpret the various regression coefficients from the model. First, report and interpret the unstandardized regression coefficient for the binary contact predictor in APA format (e.g., this should only take 1-2 sentences). [5 points]

Shoppers’ contact with an employee was associated with a 0.08 unit higher satisfaction level (*B* = 0.08, *p* = .31) compared to no contact when other independent variables are held constant. However, this association was not statistically significant (*p* = .31).

1. Report and interpret the unstandardized results for the department categorical variable in APA format. [5 points]

The automotive department (coded as 2) was associated with a 0.14 unit increase in mean satisfaction score (*B* = 0.14, *p* = .17) in comparison to the clothing department. However, this difference was not statistically significant. In contrast, the electronics department (coded as 3) was associated with a 0.21 unit increase in mean satisfaction score (*B* = 0.21, *p* < .05) in comparison to the clothing department, and this difference was statistically significant.

1. Report and interpret the unstandardized coefficient for the continuous age predictor in APA format. [5 points]

A one-unit increase in age was associated with an approximate 0.004 unit increase in mean satisfaction level (*B* = 0.0037, *p* = .17) when controlling for other independent variables. However, this association was not statistically significant (*p* = .17).

**Regression Diagnostics**

1. Multicollinearity: estimate the variance inflation factor (VIF) values for the above regression model. Paste the results below and describe if there are any apparent issues of multicollinearity. [3 points]

GVIF Df GVIF^(1/(2\*Df))

age 1.014543 1 1.007245

contact 1.009893 1 1.004934

as.factor(dept) 1.009904 2 1.002467

The VIF values for all three variables are less than 5, which indicates that there are no issues with multicollinearity.

1. Normality of the residuals: plot a histogram of the regression residuals. Paste the figure below and describe if the distribution appears approximately normal. [3 points]

A graph of a number of individuals

Description automatically generated

Yes, the distribution of the residuals appears approximately normal. Thus, the model doesn’t violate the assumptions of normality of variance.

1. Using the plot() function, check the homogeneity of variance in the residuals assumption. Paste the figure below and describe if it appears that the assumption is met. [3 points]

A diagram of a graph

Description automatically generated with medium confidence

The figure indicates that the residuals do not show any significant pattern or systematic change in variance across the fitted values. Therefore, the homogeneity of variance assumption appears to be satisfied.

***Homework #1, Part B: Data Wrangling Longitudinal Data***

The **HW\_time1** and **HW\_time2** Excel data files on Canvas will be used for Part B. Imagine these two data files contain scores on math test performance both before and after a general self-esteem intervention was carried out for those in a control condition (condition = 0) and the experimental condition (condition = 1).

1. Import both Excel files into RStudio as separate data frames.
2. You know that there were some dropouts in the study. Your mentor wants you to do analyses only on participants who remained in the study (i.e., have data at both time points). Do an appropriate merge/join that produces a merged data frame that only has participants who were in both time points (i.e., there should be no missing (NA) values anywhere in the merged data frame). [5 points]

HW\_time2 <- rename(HW\_time2, "id" = "id\_T2") # rename id column

merged\_df <- inner\_join(HW\_time1, HW\_time2, by = "id") # inner join

1. You first want to do some analyses that do not take condition into consideration. First, estimate and report the **overall means** and **standard deviations** for the math test scores at the two time points. [4 points]

Time1\_mean Time1\_sd Time2\_mean Time2\_sd

<dbl> <dbl> <dbl> <dbl>

50.0 9.48 54.6 10.7

1. Perform a paired-samples *t* test to contrast the time 1 and time 2 math scores. Report the results below in APA format. [5 points]

A paired sample t test revealed that math scores at time 2 (M = 54.6, SD = 10.7) were significantly higher than math scores at time 1 (M = 50.0, SD = 9.48), *t* (234) = -5.03, *p* < .001. This indicates a statistically significant improvement in math scores between the two time points.

1. Reshape the merged data frame from wide to long format and save to a new data frame. Each participant should have two rows of data after reshaping. Copy/paste your R syntax below. [5 points]

merged\_df <- rename(merged\_df, "math\_T1" = "math")

long\_df <- gather(merged\_df, key = "time", value = "math\_score", math\_T1, math\_T2)

1. There are somethings about the long data frame that you wish to change for presentation/analysis purposes. **First**, because the condition assignment for participants was the same across the two time points, you only want one condition column (i.e., remove one of the condition columns). **Second**, you want numerical values in the key column that represents time point (i.e., have the key for time 1 equal 1 and the key for time 2 equal 2). **Finally**, you want to order/arrange the data frame based on participant ID and time point. That is, you want the first two rows to be from participant 1, the next two rows from participant 2, etc. Make sure all these changes are applied to the long data frame. You should be able to view the data frame to ensure the changes were applied correctly. [6 points]

long\_df <- long\_df %>%

select(!condition\_T2) %>%

mutate(time = if\_else(time == "math\_T1", 1, 2)) %>%

arrange(id)

1. Output the means and standard deviations for the math scores at the two time points for each condition by grouping the data by condition and time point and summarizing the math scores scores. Report the values below. Do these descriptive statistics suggest that the paired-samples t test results above is being driven by one of the conditions? Explain your reasoning. [5 points]

condition time Mean SD

<dbl> <dbl> <dbl> <dbl>

1 0 1 49.5 9.81

2 0 2 49.8 9.85

3 1 1 50.6 9.15

4 1 2 59.3 9.46

The descriptive statistics suggest that the results of the paired-samples t-test are primarily driven by the experimental condition (coded as 1). Specifically, in the experimental condition, there was a difference of approximately 9 points in math scores between the two time points. In contrast, the control condition showed a minimal difference of only 0.3 points. This indicates that the significant difference observed in the paired-samples t-test is largely attributable to the experimental condition alone.

***Homework #1, Part C: Null Multilevel Models***

Part C uses the **cognitive.dta** Stata data file on Canvas. These data focus on various cognitive test scores in students nested within schools.

1. Import the cognitive data file into RStudio. [1 point]
2. Execute a null multilevel model (using the lmer() function) with the school as the group/cluster variable and the arithmetic math score as the dependent variable. The arithmetic math scores derive from a test that had 17 questions (i.e., scores could range from 0 to 17). Make sure that maximum likelihood (ML) is used as the estimator. Paste your R syntax below. [10 points]

null\_model <- lmer(arithmetic ~ (1|schoolid), REML = FALSE, data = df)

summary(null\_model)

1. Report the following fit statistics from the null model: [4 points]
   1. AIC: 10316.1
   2. BIC: 10333.7
   3. Log-likelihood: -5155.1
   4. -2LL: 10310.1
2. Report the following sample sizes from the data: [3 points]
   1. Number of students: 2598
   2. Number of schools: 12
   3. Average number of students per school: 216.5
3. Report the following variance components from the model output: [2 points]
   1. Between-group variance: 0.1409
   2. Within-group variance: 3.0648
4. Calculate and report the intraclass correlation coefficient (ICC) and the significance (*p*-value) associated with the random intercept. Based on these values, decide if multilevel modeling does seem necessary for this data and explain your reasoning. [5 points]

The ICC value of 0.045 indicates that the grouping variable has a small clustering effect, accounting for approximately 4.5% of the total variance in the outcome. While this suggests that the clustering effect is modest (and may not require multilevel modeling), the random effect is statistically significant (p < .001), indicating that the grouping variable does have a meaningful impact on the outcome. Given the significance of the random effect, multilevel modeling is recommended.

1. Report the unstandardized coefficient for the intercept fixed effect and interpret the value on the arithmetic test score range. [5 points]

The unstandardized coefficient for the intercept fixed effect is *B* = 7.822, *p* < .001. This indicates that the average (grand mean) arithmetic test score across all schools is 7.822. This falls within the lower to middle range of the possible scores (0 to 17)