Comprehensive Homework 1

PSYCH 610, 2023 - 24

Due Oct 11th, 1:30 PM. 100 points total.

**Instructions**

Welcome to your first comprehensive homework! Remember, you are allowed to refer to prior lab scripts, homework assignments, books/articles, and the internet (including ChatGPT). **You are not allowed to consult with other students**, including more advanced students in your program. Use RMarkdown as you do for homework assignments. A few things to note:

1. **Submit your files anonymously.** Submit both .Rmd and .html files. Instead of using the normal Lastname\_HWXX format, use the **last 4 digits of your student ID number** (e.g., “3180\_comp\_01.Rmd”). **Include NO identifying information in your files!!!**
2. You can email us asking clarification questions, but we will not be as instructive as we would be on a regular homework assignment. Do not post questions to Slack channels or where other students can see.
3. The number of points each question is worth appears in parentheses next to the question, for a total of 100 points. 2 of the 100 points will be awarded if the correct files were submitted with appropriate file names (1) and correct formatting (1).

**Formatting**

To obtain 2 out of 2 points, you must adhere to the following:

1. Label your answers in RMarkdown according to *sections*, *question numbers*, and *question letters*
   1. Use RMarkdown heading level 1 for section headings (e.g., “Conceptual questions”, “Data Analysis”, etc.).
   2. Use RMarkdown heading level 2 for question numbers.
   3. Use RMarkdown heading level 3 if there are question letters.
2. Always start your answer on a new line following any question number and letter, so your response is not formatted like a heading.
3. Do NOT repeat the question prompt in RMarkdown; question numbers and letters are sufficient for the TAs to grade the homework.
4. Follow the course style guide and norms for naming variables and models.
5. Knit your file to a .html document. When you’re done knitting, check the output .html file to ensure that your answers are readable and formatted correctly in the .html format.

If you have questions about this assignment, **email** or **direct message LiChen Dong** in our class Slack channel ([ldong54@wisc.edu](mailto:ldong54@wisc.edu)), or **drop by his office hours**.

**Conceptual Questions (23 pts)**

1. Shourya, a graduate student, was curious whether depression would influence individuals’ physical health. To test his hypothesis, Shourya recruited five other graduate students from his program. To manipulate depression, he asked them to choose between two YouTube videos: one was a comedy show and the other was a tragic story. The five participants received a link to the video they chose and were asked to watch the video whenever they had spare time. To measure physical health, Shourya had the participants answer a question (“From 1 to 5, how would you rate your physical health?”) before and after watching the video and used the difference score as his main outcome of interest. Shourya then used the type of video that participants had chosen (operationalization of “depression”) to predict the difference score of physical health.
   1. Do you think Shourya has a reliable measure of physical health? Why? (1)
   2. What are the four types of validity? Is Shourya’s study satisfactory in terms of these four types of validity? Why or why not? (4)
   3. List two changes that improve Shourya’s experiment in terms of reliability and/or validity. Explain how each change will help reliability and/or validity. (2)
2. Answer the following questions about sampling distribution:
   1. Suppose UW-Madison students have a mean age of 20.5 with a standard deviation of 1.5. Assuming that student age follows a normal distribution and Erin, who is a UW-Madison student, has a z-score of 1.0, how old is Erin (in years)? (1)
   2. What is the definition of standard error (in relation to a sampling distribution)? (1)
   3. Why do we usually use the t-distribution, not the z-distribution, to test hypotheses? (1)
3. A basic/mean-only model can be written as: Yi = b0 + ei
   1. Other than the sum of squared errors (SSE), list two alternative ways to aggregate separate error values (ei) into a summary measure of total error (see textbook). (2)
   2. If we use the sample *median* instead of the sample mean as our parameter estimate for b0, what summary measure of total error from question 3a will be minimized? (1)
   3. What is mean squared error (MSE) as defined in the textbook? In the basic/mean-only model, what another special name do we give to MSE? (1)
4. Consider the comparison of two regression models that tests the effect of a predictor X on Y

Model C: Yi = b0 + ei where SSE = 1000

Model A: Yi = b0 + b1\*Xi + ei where SSE = 750

* 1. Calculate the η2 (partial-eta2) for predictor X. Explain what this number means. (2)
  2. How would mean-centering the predictor X change our parameter estimate for b1? Be precise: Would it increase, decrease, stay the same, or we don’t know? (1)
  3. Assuming a sample size of 27, calculate F statistic for predictor X (show your work). (1)
  4. Consider the formula for F statistic, what does the numerator and denominator of F statistic each represent, in terms of errors and additional parameters? (2)

1. Sara wanted to examine if it is possible to predict weekly alcohol consumption (measured as “number of drinks”) from participant self-reported mood swings. After testing a simple regression model, Sara found that the effect of mood swings was statistically significant, b1 = 0.78, SE = 0.26, t(100) = 3.01. Help Sara calculate the 95% confidence interval (CI) for b1. Make sure to show your work and feel free to do the math using R code. What does the 95% CI for b1 tell Sara about the effect of mood swings? (A table with critical t-values can be found at https://connect.springerpub.com/content/book/978-0-8261-9825-9/back-matter/bmatter3). (3)

**Data Analysis**

In this comprehensive homework, you will be working with *quasi-real (anonymized and randomized)* data, adapted from the Public-Use Data Sets of the National Longitudinal Study of Adolescent to Adult Health (Add Health). See <https://addhealth.cpc.unc.edu/> for more information.

Cognitive ability and personality are two well-established aspects of individual differences, both of which are relatively stable across adulthood. However, it remains unclear how cognitive ability and personality influence the development of one another over the lifespan. A group of researchers here at UW-Madison want to investigate whether cognitive ability assessed in adolescence would predict *imagination*, one of the factors in the Five-Factor Model of personality, measured in adulthood. They are also interested in how *agreeableness*, another factor of personality, may be influenced by supportive parenting, which is an important aspect of early family environment that has substantial influence on the development of many behavioral traits. The researchers want to test their hypotheses using data from Add Health and have asked for your help in preparing the data and conducting some statistical analyses.

The sample consists of 3,093 participants. Cognitive ability was assessed when the participants were adolescents, using a short version of the Peabody Picture Vocabulary Test (AH\_PVT). Supportive parenting was measured in adolescence and recorded as a dichotomous variable (Sup\_Parent), reflecting either high or low supportive parenting. Lastly, a set of questions from the Mini-IPIP scale was used to measure the personality traits of the Five-Factor Model when the participants were adults. You selected the items from that scale which are pertinent to *agreeableness* (Agr\_1 -Agr\_4)and *imagination* (Ima\_1 -Ima\_4). You have not evaluated the psychometric properties (i.e., reliability) of these items yet.

The researchers ask you to test the following hypotheses:

1. Cognitive ability is significantly associated with imagination.
2. Supportive parenting is significantly associated with agreeableness.

Codebook:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Coding Scheme | Description | Forward (+)/ Reverse (-) |
| AID | 10316654 - 99886999 | Participant ID number | NA |
| SCID | 001 - 999 | Participant school ID number | NA |
| AH\_PVT | 50 - 146 | Cognitive ability score | NA |
| Sup\_Parent | 1 = High, 2 = Low | Supportive parenting (dichotomous) | NA |
| Parent\_Div | “Married”, “Divorced” | Parent marital status (dichotomous) | NA |
| Agr\_1 | 1 (Strongly Disagree)  to  5 (Strongly Agree) | “I sympathize with others' feelings.” | + |
| Agr\_2 | “I feel others' emotions.” | + |
| Agr\_3 | “I am not really interested in others.” | - |
| Agr\_4 | “I am not interested in other people's problems.” | - |
| Ima\_1 | “I have a vivid imagination.” | + |
| Ima\_2 | “I have difficulty understanding abstract ideas.” | - |
| Ima\_3 | “I am not interested in abstract ideas.” | - |
| Ima\_4 | “I do not have a good imagination.” | - |

1. Read in and inspect the data (Total 7 pts)
   1. Which variables are predictors (IVs) and which ones are outcome variables (DVs)? (1)
   2. Read in the data and save them under a name consistent with the class style guide. (1)
   3. Using a single line of code, format the variable names to follow the naming convention for our class (hint: we use a particular “case” for variable names). (1)
   4. Use at least two functions that you learned in lab to briefly inspect the data. Are there any missing values (NAs)? How do you know? (2)
   5. Use one function to obtain descriptive statistics for all variables in the dataframe. (1)
   6. There are two nuisance variables in the dataset which you will not use for your analysis: AID and SCID. Remove the two variables and run a command that shows the names of the variables that are left. (1)
2. Assess the reliability of the scales and create composite scores (Total 15 pts)
   1. Check the reliability of the agreeableness scale. What is Cronbach’s alpha of this scale? Do you consider this scale reliable? Why? (2)
   2. For each item on the agreeableness scale, write out Cronbach’s alpha of the scale if that item is dropped. (1)
   3. Check the reliability of the imagination scale. What is Cronbach’s alpha of this scale? Do you consider this scale reliable? Why? (2)
   4. Using a single function, create a simple correlation table with all items of the imagination scale that will show up in R console. (1)
   5. These items are in fact from a pre-established and validated personality scale: Mini-IPIP, a 20-item short form of the 50-item International Personality Item Pool—Five-Factor Model measure. Look at the Mini-IPIP website <https://ipip.ori.org/MiniIPIPKey.htm>, in particular the expected psychometric properties (i.e., reliability) of each scale. Now, what do you think about the reliability of the agreeableness and imagination scales in our sample? Is reliability the higher the better? Reflect on measurement reliability in real-world settings. Write down your thoughts in a few sentences (2).
   6. Create a composite score of agreeableness using items Agr\_1 – Agr\_4 for individuals who answered *at least three items* on the agreeableness scale. (2)
   7. Create a composite score of imagination using items Ima\_1 – Ima\_4 for individuals who answered *at least three items* on the imagination scale. (2)
   8. Create centered versions of *all* predictor variables. [Note that you could answer all of the questions below without centering your predictors, but because the instructor team wants you to practice everything you have learned, we are asking you here to center the predictor variables and to use the centered variables by default. (3)
3. Show descriptive statistics and “quick-and-dirty” plots of the data (Total 6 pts)
   1. Create a “quick-and-dirty” histogram for the cognitive ability variable. (1)
   2. Create a “quick-and-dirty” scatterplot representing the relationship between cognitive ability and the imagination composite score. Include a regression line in your plot. (1)
   3. Create a *character* version of the supportive parenting variable that has meaningful text labels/values for the high and low supportive parenting groups. (1)
   4. With the *character* variable that you just created, use a single function to find out the number of individuals in high vs. low supportive parenting group. (1)

*Bonus point*: within the function above, use one argument to also show the number of individuals who have NAs on the supportive parenting variable. (+1)

* 1. With the *character* variable that you just created in Question 8c, use a single function to obtain both the means and standard deviations of the agreeableness composite score broken down by supportive parenting. (1)
  2. Using a single function, create a simple correlation table with ah\_pvt, sup\_parent\_c, the agreeableness composite score, and the imagination composite score. (1)

1. Review the null model and the mean-only model (Total 5 pts)
   1. Calculate the SSE for the null model that predicts the imagination composite score using an a priori value of 0. Do so using a step-by-step brute force method in R. (1)
   2. Calculate the SSE for the mean-only model that predicts the imagination composite score using the sample mean. Do so using a step-by-step brute force method in R. (1)
   3. In lab, we introduced how to estimate and compare the null model and the mean-only model using lm() and anova(). Compare the two models above using that method, *without* using summary(). What is the F statistic, degrees of freedom, and p-value? (1)
   4. What is the conclusion from the model comparison above? (1)
   5. Fill in the blank: the model comparison above is also called an “\_\_\_\_\_\_\_\_\_ t-test.” (1)
2. Statistical analysis: imagination (Total 16 pts)
   1. Model\_1. Fit a model to test the hypothesis that cognitive ability is associated with imagination, using the mean-centered version of the predictor. (1)
   2. Report F statistic, degrees of freedom, and p-value for b1, i.e., the regression coefficient associated with the predictor in Model\_1. Briefly interpret the corresponding p-value. (2)
   3. Use code to obtain the effect size of the relationship in terms of η2. Also provide its interpretation in a sentence to describe the effect of cognitive ability. (2)
   4. Use code to obtain the 95% confidence interval around the intercept and the slope. (1)
   5. Interpret the parameter estimate for intercept (b0). (1)
   6. Interpret the parameter estimate for slope (b1). (1)
   7. How many parameters are estimated in Model\_1? What are they? (1)
   8. Model\_2. Fit another model where you use the *uncentered* version of the cognitive ability variable to predict imagination. Did the parameter estimate for intercept (b0) change? Did the parameter estimate for the slope (b1) change? Why or why not? (3)
   9. Create a publication-quality graph that depicts Model\_2. Make sure to include the appropriate error band, a title, axis labels, and points representing the raw data. (4)
3. Statistical analysis: agreeableness (Total 16 pts)
   1. Model\_3. Fit a model to test the hypothesis that supportive parenting is associated with agreeableness, using the recoded -0.5/ 0.5 version of the predictor. (1)
   2. Report F statistic, degrees of freedom, and p-value for b1, i.e., the regression coefficient associated with the predictor in Model\_1. Briefly interpret the corresponding p-value. (2)
   3. Use code to obtain the effect size of the relationship in terms of η2. Also provide its interpretation in a sentence to describe the effect of supportive parenting. (2)
   4. Use code to obtain the 95% confidence interval around the intercept and the slope. (1)
   5. Interpret the parameter estimate for intercept (b0). (1)
   6. Interpret the parameter estimate for slope (b1). (1)
   7. Create a publication-quality graph that depicts Model\_3. Make sure to include the appropriate error bars, a title, axis labels, and horizontally-jittered raw-data points. (4)
   8. Model\_4. Recode the supportive parenting variable such that individuals with high supportive parenting have a value of 1, and those with low supportive parenting have a value of -1. Fit another model where you use *this recoded version (-1 vs. 1)* of the supportive parenting variable to predict agreeableness. Did the model estimate for intercept (b0) change? Did the slope (b1) change? Why or why not? (4)
4. Write up your results (Total 5 pts)
   1. Write a concise paragraph for the Results section in a paper, where you set up the context of the study [which admittedly would usually be in the Introduction or in the Methods section], outline your specific hypotheses, describe the models you fit, discuss the results of the analyses, and interpret your results in non-jargony terms so that a non-specialist can understand your interpretation. (5)
5. Conduct a post-hoc explorative analysis (Total 4 pts)
   1. Create two new dataframes by subsetting the data according to parental divorce status (e.g., d\_married & d\_divorced), as indicated by the variable parent\_div. For each new dataframe, fit a model in which you use the -0.5/ 0.5 version of the supportive parenting variable to predict the agreeableness composite score. Look at the summaries of the two models. Does the effect of supportive parenting on agreeableness seem rather similar or wildly different across families with divorced vs. non-divorced parents? (4)
6. How long did this assignment take you to complete? (1)

**Congratulations on finishing Comprehensive Homework 1!**