

Technical Writing Artifact: SPSS Statistical Analysis Instructional Guide

Author: Dr. Maureen K. Johnson

This document is an example of technical instructional material created for graduate students. It demonstrates step-by-step procedural writing, statistical explanation, and clear communication of analytical workflows.

AHS 604: SP 2020

PRACTICE ASSIGNMENT 3 INSTRUCTIONS

INSTRUCTIONS: This document provides students with practice analyzing data in SPSS and reporting findings in an APA formatted table. Your work for the tasks discussed in this document will not be graded. However, you will need to apply similar steps to Assignment 3 as in Practice Assignment 3; for this reason, it is strongly recommended you complete Practice Assignment 3. Responses to Practice Assignment 3 will be posted so that you will be able to check your responses. Please let me know if you have any questions or need clarification concerning the responses to Practice Assignment 3. I strongly recommend watching the tutorial videos in the “Helpful Videos” folder (all videos were made by TheRMUoHP Biostatistics Resource Channel. These instructions also contain links to other helpful videos. Complete the following tasks:

- 1. RETRIEVE THE DATA SET AND OPEN IN SPSS.** The SPSS data set you will need to retrieve is called “**AHS 604 Data Set for Practice Assignments 1, 2, and 3 SP 2020.sav**”. Refer to the screenshots from the Instructions for Practice Assignment 1 to assist you with retrieving SPSS data sets for the class.
- 2. CALCULATING MULTIPLE LINEAR REGRESSION: PREDICTING ALCOHOL INTAKE WITH AGE, GPA, EXERCISE TIME, AND WEEKLY COFFEE CONSUMPTION.**
 - a. RESEARCH QUESTION:** For this problem, our research question is “Are age, GPA, weekly exercise time (in hours), and weekly coffee consumption significant predictors of alcohol intake?” In other words, we are testing to see if there is sufficient evidence to suggest age, GPA, exercise time, and weekly coffee consumption are significant predictors of alcohol intake. We will test the distribution of all continuous independent variables of age,

GPA, exercise time, and weekly coffee consumption as well as our dependent variable of alcohol for normality. Then, we will generate descriptive statistics (e.g., mean, standard deviation. Then, we will determine intercorrelations between variables. Then, we will look at our test statistic, which is our beta (β) coefficient as well as our level of significance. Then, we will determine our effect size. Complete the following:

- b. **TEST VARIABLES FOR NORMALITY.** Linear regression assumes the data for the variables have normal distributions. One way to determine if our distribution is normal is to determine the skewness and kurtosis of the distribution (as we did with Practice Assignments 1 and 2).
 - i. Click on “Analyze”, then “Descriptive Statistics”, then “Frequencies”.
 - ii. The “Frequencies” window will now be opened. Then, click on the following variables and move them into the “Variable(s)” field on the right: “Age”, “Hours_Exer”, “Avg_Drinks”, “GPA”, and “Coffee_Week”.
 - iii. Click on the “Statistics” button on the right. The “Frequencies: Statistics” window will appear.
 - iv. Under “Central Tendency”, check the “Mean”, “Median”, and “Mode” boxes. Under “Dispersion”, check “Standard deviation”. Under “Distribution”, check “Skewness” and “Kurtosis”. Then click on the “Continue” button at the bottom of the window.
 - v. Now, you should be back in the “Frequencies” window. Click on the “Charts” button on the right hand side of the window.
 - vi. The “Frequencies: Charts” window now appears. Under “Chart Type” check on “Histograms”. Then, under “Histograms:”, check “Show normal curve on histogram”. Then click on the “Continue” button at the bottom of the window.
 - vii. Now, you are back in the “Frequencies” window. Be sure to check the “Display frequency tables” box at the bottom of the window. Then click on the “OK” button.
 - viii. You will see an Output document which will show frequency tables (with all requested descriptive statistics). It will also show a histogram (the connecting bars) showing the distribution of participants by age, hours of weekly exercise, average number of alcoholic drinks, GPA, and weekly coffee intake. You also see the outline of the normal (bell-shaped) curve superimposed over the histogram. You will use the skewness and kurtosis statistics to determine whether or not the data for each of our variables are normally distributed. You will address this in your “Discussion of the Results” section in Part d.

ix. In addition to the skewness statistic for all of the variables in the linear regression model, run the Kolmogorov-Smirnov and Shapiro-Wilks tests for normality on the dependent variable of alcohol intake as follows:

1. Click on “Analyze”, then “Descriptive Statistics”, then “Explore”.
2. The “Explore” window will now be opened. Then, click on the “Avg_Drinks” variable and then the right-pointing arrow to move it into the “Dependent List” field on the right. Also, at the bottom of the “Explore” window, under Display, check the button for “Both.”
3. While still in the “Explore” window, Click on the “Statistics” button on the right hand side of the window. When the “Statistics” window opens, make sure the box next to Descriptives is checked and the Confidence Interval for the Mean is indicated as 95%. Then, click on the “Continue” button.
4. When you are back in the “Explore” window, click on the “Plots” button. When the “Plots” window opens, check the button next to “Factor levels together” (if this button is not already checked), check the box next to “Stem-and-leaf”, and check the box next to “Normality plots with tests”. Then, click on the “Continue” button.
5. Now, you should be back in the “Explore” window. Click on the “OK” button, and SPSS will generate the results of the the Kolmogorov-Smirnov and Shapiro-Wilks tests.

c. CALCULATING THE TEST STATISTIC

- i. For this problem, our research question is “Are age, GPA, weekly exercise time (in hours), and coffee intake significant predictors of hours of weekly alcohol intake?” To answer this question, we will need to generate descriptive statistics, correlations, and the results of the multiple linear regression analysis.
- ii. Click on “Analyze”, then “Regression”, then “Linear.” Then, the “Linear Regression” window appears. On the left side of the window, all of the variables in your data set are listed.
 1. **SELECTING THE DEPENDENT VARIABLE:** Highlight “Avg_Drinks” and then click on the arrow next to the ‘Dependent’ field. Now, this variable should be listed in the “Dependent” field.
 2. **SELECTING THE INDEPENDENT VARIABLES:** Highlight each of the variables “Age”, “GPA”, “Hours_Exer”, and “Coffee_Week” and move each of them into the “Independent(s)” field.

3. **ENTER METHOD OF REGRESSION:** Next to Method, select “Enter”.
 4. **STATISTICS WINDOW:** Click on the “Statistics” button. In the “Statistics” window, complete the following:
 - a. Under “Regression Coefficients”, check “Estimates”, “Confidence Intervals”, “Model fit”, “Descriptives”, “Part and partial correlations”, and “Collinearity diagnostics”.
 - b. Then, click the “Continue” button.
 5. **GENERATING PLOTS:** Once back in the “Linear Regression” window, click on the “Plots” button. In the “Linear Regression: Plots” window, complete the following:
 - a. Select “*ZRESID” and move it to the “Y:” field.
 - b. Select “*ZPRED” and move it to the “X:” field.
 - c. Under “Standardized Residual Plots” check “Histogram” and “Normal probability plot”.
 - d. Click on the “Continue” button.
 6. **GENERATING SPSS OUTPUT:** Click on the “OK” button at the bottom of the Linear Regression window. Check to make sure the results of your linear regression analysis appear in your output sheet. Export your output to a Rich Text File (as we did with Practice Assignments 1 and 2). Be sure to select “Shrink width to fit page margins” so that all tables are exported to RTF/Word format.
- d. **TABLES AND DISCUSSION.** For this section, you will need the SPSS output you generated in the “Calculating the Test Statistic” section. You will also need to refer to the Table Template APA Format (which is in Word format) posted in Blackboard. While the SPSS output provides you with tables, these tables are not in the format which is used to report findings in a research paper. Thus, you can use the appropriate table template to report your findings in APA (6th edition) format.
- i. **GENERATING THE TABLES:**
 1. Select Tables 8 and 9 in the Table Template APA Format document and copy them into the imported Word document (Rich Text File) with your SPSS output from the “Calculating the Test Statistic” section. Using the template for Table 8, you will enter the means and standard deviations for all variables in the analysis as well as correlations between the dependent variable and independent variables. You will also enter correlations between (or among) the independent variables.
 2. For summarizing the results of the regression analysis, use the template for Table 9. Here, you will report the unstandardized coefficient (*B*), standard error of the

- standardized coefficient (*SE B*), the standardized beta coefficient (β), the *t*-statistic for each independent variable, and the *p*-value.
3. For a footnote beneath the regression summary table, you can report the results of R^2 , which is the effect size statistic for the regression (include the sample size and the *p*-value). This can also be included in the discussion of the findings.
 4. Be sure to give the tables and appropriate headings and enter the variables in the format shown in the table template. However, to insert additional rows to this table (if needed), right-click on the last line of data, and a menu will appear. Then, click on “Insert” and then “Insert Rows Below”. Insert new rows until you have a sufficient number of rows for entering your data. If you ever need to delete rows, you can right-click on the row you need to delete, select “Delete Cells”, and then select “Delete entire row”.
- ii. **DISCUSSION OF RESULTS:** Once you have your data entered in the table, discuss your findings. To summarize your findings, discuss the mean and standard deviation of the distribution of each variable for the sample (the discussion must be in paragraph format rather than in Q & A format). When reporting our findings, we need to answer the research question we had raised. We will also discuss the following:
1. Our research hypothesis (H_1).
 2. The reason we used multiple linear regression for our statistical test (refer to the independent and dependent variables as well as the number of groups we have).
 3. The skewness (symmetry) and kurtosis (steepness or “peakedness”) of our distributions for all of the variables (independent and dependent) in the linear regression model.
 4. The results of the Kolmogorov-Smirnov and Shapiro-Wilks tests for normality on the dependent variable of alcohol intake.
 5. The means and standard deviation of all variables entered into the regression analysis.
 6. Statistically significant correlations (r) between variables (including the *p*-value).
 7. The standardized coefficient (β), the *t*-statistic, and the *p*-value for any statistically significant predictor.
 8. The coefficient of determination (R^2), which determines the effect size of the relationship between the predictor (independent) variables and the outcome (dependent) variable accompanied by other relevant test statistics (*F* and *p*-value) Complete the following:

- a. To interpret effect sizes, review slide 28 of Dr. Karabi Nandy's presentation "Understanding and Quantifying Effect Sizes", which can be viewed at <http://docplayer.net/4361544-Understanding-and-quantifying-effect-sizes.html> .
- b. Referring to slide 28 of Dr. Nandy's presentation , describe the strength of the percentage of variation in number of drinks per week accounted for by variation in the predictor variables (the constant, age, GPA, hours of exercise per week, and weekly coffee intake)?

3. CALCULATING LOGISTIC REGRESSION: PREDICTING CURRENT SMOKER STATUS WITH MULTIPLE INDEPENDENT VARIABLES (ORDINAL AND CONTINUOUS).

- a. **RESEARCH QUESTION.** For this problem, our research question is "Are year in college, weekly exercise time (in hours), weekly alcohol consumption, and weekly coffee consumption significant predictors of current smoking status?" In other words, we are testing to see if there is sufficient evidence to suggest year in college, weekly exercise time (in hours), and weekly alcohol consumption, and weekly coffee consumption are significant predictors of current smoking status. Although descriptive statistics are generally reported for a logistic regression, we will focus on our test statistic, which is our beta (β) coefficient as well as our level of significance. However, if any of our β 's are statistically significant, we will determine our effect size, which can be done with an odds ratio (OR) and on the SPSS output is designated as "Exp (B)". Complete the following:
- b. **CALCULATING THE TEST STATISTIC**
 - i. Click on "Analyze", then "Regression", then "Binary Logistic." Then, the "Logistic Regression" window appears. On the left side of the window, all of the variables in your data set are listed.
 - ii. **SELECTING THE DEPENDENT VARIABLE:** Highlight the "Curr_Smoke" variable and then click on the arrow next to the "Dependent" field. Now, this variable should be listed in the "Dependent" field.
 - iii. **SELECTING THE COVARIATES (INDEPENDENT VARIABLES):** Select each of the variables of Avg_Drinks , Coffee_Week , Sch_Year, and Hours_Exer listed on the left hand side of the "Logistic Regression" window and click on the arrow to the left of the "Covariate(s)" field. Move each of these variables until these four variables are listed in the "Covariate(s)" field.

- iv. **ENTERING METHOD:** For the “Method” menu underneath the “Covariates” field, be sure that “Enter” is indicated.
- v. **CATEGORICAL VARIABLES:** Click on the “Categorical” button in the upper right hand corner of the “Logistic Regression” window. We have one categorical variable (Sch_Year) which needs to be indicated as a “Categorical Covariate”. If this variable is not already in the “Categorical Covariates” box, highlight it and click on the arrow pointing to the right next to the Categorical Covariates” box. Now, you should see “Sch_Year” in the “Categorical Covariates” box. In the “Categorical Covariates” box, highlight “Sch_Year”. Then, under “Change Contrast” pull down the Contrast menu and select “Indicator”. Next to “Reference Category” check the button next to “First”, and then click on the “Change” button. This means that for “Sch_Year”, we are making the first listed group “Freshmen” our reference group to which we are comparing Juniors, Seniors and Sophomores in terms of their ability to predict current smoking behavior. Click the “Continue” button at the bottom of the window.
- vi. **CLASSIFICATION ACCURACY.** In the “Logistic Regression” window, click on the “Save” button. When the “Logistic Regression: Save” window opens, check the box for “Group membership” underneath “Predicted Values”. Then, click the “Continue” button. We are doing this because we want to see if our logistic regression model (with the predictor variables) more accurately classifies individuals as a current smoker than the null model (which only contains the constant and none of the predictor variables).
- vii. **ADDING OPTIONS:** Back in the “Logistic Regression” window, click on the “Options” button on the upper right-hand side of the window. Now, you are in the window “Logistic Regression: Options”. In the section “Statistics and Plots”, check on the boxes for “Hosmer-Lemeshow test” and for “CI for exp(β)”. In the “Display” section, be sure “At each step” is marked. At the bottom of the window, be to check the box for “Include constant in model.” Then, click on the “Continue” button. For the Hosmer-Lemeshow test, we will actually hope to see a significance level greater than 0.05.
- viii. Once you are back in the “Logistic Regression” window, click on the “OK” button at the bottom of the window. Check to make sure the results of your logistic regression analysis appear in your output sheet.
- ix. Export your output to Word/Rich Text File (as you did with Practice Assignment 2).

c. IMPORTANT INFORMATION CONCERNING THE CATEGORICAL VARIABLES STATE OF SCHOOL YEAR:

- i. Refer to the “Categorical Variables Codings” table of the SPSS output. For “Gender”, we will be comparing the ability of being female (represented by “Gender(1)” in the “Variables in the Equation” table) to that of being male (which is the last gender category and is the reference).
 - ii. Refer to the “Categorical Variables Codings” table of the SPSS output. For school year (Sch_Year) you see that the “Freshman” category is first. Recall that in Part 6. b. v., we made the first category of “Sch_Year” our reference group. Thus, the “Freshman” category, which is listed first, is our reference group for this analysis. The remaining categories are alphabetically arranged. Therefore, each of three different levels of “Sch_Year” (juniors, seniors, and sophomores) of year in college is represented by the variables Sch_Year(1), Sch_Year(2), and Sch_Year(3), respectively. Referring to the table with the heading “Variables in the Equation” in your output for the Binary Logistic Regression analysis, you see that each of these levels is compared to college freshmen in terms of its ability to predict whether or not a student is a current smoker.
- d. TABLES AND DISCUSSION. For this section, you will need your SPSS output from Question 5 Parts b and c. You will also need to refer to the Table Template APA Format (which is in Word format) posted in Blackboard. While the SPSS output provides you with tables, these tables are not in the format which is used to report findings in a research paper. Thus, you can use the appropriate table template to report your findings in APA (6th edition) format.
 - i. GENERATING THE TABLES:
 1. Select Table 10 in the Table Template APA Format document and copy it into the imported Word document (Rich Text File) with your SPSS output from the Part b “Calculating the Test Statistic” section. Using the template for Table 10, you will enter the beta coefficient (B), the standard error (S.E.), odds ratio (ExpB), 95% confidence interval for the odds ratio ((5% CI for ExpB), the Wald Statistic (Wald), and the p-value (Sig.).
 2. For a footnote beneath the regression summary table, you can report the results of R^2 , which is the effect size statistic for the regression (include the sample size and the p-value). This can also be included in the discussion of the findings.
 3. Be sure to give the tables and appropriate headings and enter the variables in the format shown in the table template. However, to insert additional rows to this table (if needed), right-click on the last line of data, and a menu will appear. Then, click on “Insert” and then “Insert Rows Below”. Insert new rows until you have a sufficient number of rows for

entering your data. If you ever need to delete rows, you can right-click on the row you need to delete, select “Delete Cells”, and then select “Delete entire row”.

- e. In addition to the Logistic Regression Analysis in SPSS video in the Helpful Videos folder in Blackboard, you may wish to watch the video “Multiple Logistic Regression in SPSS” by Practical Applications of Statistics in SPSS at <https://youtu.be/fwxFz32L-AE>. Then, discuss the following:
 - i. What is the overall fit of the model? Be sure to discuss this in terms of:
 1. Omnibus Test of Model Coefficients (report the χ^2 for the df and N-size as well as the p-value). How well does our logistic regression model fit the data? If it fits the data well, the p-value (Sig.) will be less than .05.
 2. Pseudo R square (Report the Nagelkerke R Square) as a percentage (multiplying by 100%).
 3. Hosmer and Lemeshow Test.
 - ii. Classification accuracy. How accurate is our model in predicting whether or not a student is a current smoker, especially compared to the null model (the model with only the constant and no predictor variables)? Look for the percentages correct.
 - iii. Discussion of entered predictor variables. Referring to the statistics from the “Variables in the Equation” table in the SPSS output discuss if any of the specific variables entered into the model are statistically significant predictors of being a smoker. Justify your response by referring to the Wald statistic (which is a Chi-square statistic) and the p-value. For each statistically significant predictor variable, indicate the odds ratio. If you have no significant predictor variables, indicate this is the reason you would not report odds ratios.