**Data Science for Social Scientists**

Psyc 546, Spring 2023

Homework Assignment 8

**Due Date**: March 29th (by 8:15 PM)

**Reminder**: See the assigned readings, Canvas resources, and the lecture slides for a tutorial on how to use R to perform the various functions included in the homework assignment below. **Once completed, you should submit a completed version of this document and your final R script file to the Homework 8 – Submission Portal on Canvas**.

Your submitted R script file should contain code to answer the questions below when code is necessary. Please use comments (e.g., #Question 1) to label the code for each question. **For answers where you need to report the statistical results in APA format, use the link posted to the Week 9 module on Canvas (or other online APA reporting tutorials) to help guide your answers.**

1. The **hourlywagedata.sav** data file on Canvas contains hourly wage information for a sample of nurses. Perform an independent samples t test to see if hourly wage differs between hospital and office nurses. For this question, make sure to assume equal variances. Report your answer below in APA format. Make sure to include all relevant test statistics, the statistical conclusion of the test, a description of which nurse group was higher or lower in hourly wage, and Cohen’s *d* to quantify the effect size. [1 point]

* Hourly wages were significantly higher for hospital nurses (M = 20.68) than office nurses (M = 18.69), *t*(2909) = 12.99, *p* < .001, *d* = 0.51. The effect size was medium.
* **Nurses who worked in the office (*M* = 18.69) had significantly lower hourly wages compared to nurses who worked in the hospital (*M* = 20.68). This difference was statistically significant *t*(2909) = 12.99, *p* < .001, *d* = 0.51 (medium effect).**

1. The **bankloan.sav** data file on Canvas contains information about a sample of mortgage applicants to a national bank. The bank has come up with two different models to predict the probability that an applicant will default on their mortgage (stored as variables preddef1 and preddef2). Imagine you are tasked with seeing if the two models produce significant different default rates from one another. Perform a paired samples t test to see if the predicted default rates significantly differ between the two models. Report your answer below in APA format, including all relevant test statistics and the statistical conclusion of the test. No need to report an effect size for this question. [1 point]

* The default rates produced by the two models were not significantly different from one another, *t*(849) = 0.27, *p* = 0.786.

1. The **satisfaction.sav** data file on Canvas contains data from a sample of consumers who visited a regional retail store. Satisfaction ratings were assessed for the price of the items (price), the variety of items (numitems), the organization (org), service (service), item quality (quality), and overall satisfaction (overall). Perform a correlation analysis on these satisfaction ratings using the cor() function. Because the satisfaction items are on an ordinal scale, you know that the items violate the normality assumption of the Pearson’s correlational test. As a result, make sure to perform a Spearman’s correlational test. Paste the correlation matrix/table that is printed to the R console below. You do not need to format the table in APA format or summarize any of the results, simply copy/paste the table (or a screenshot of it) below. Your correlational table should only include the above variables. That is, you will not get the point if your correlation table includes every single variable in the data file. [1 point]

Text

Description automatically generated

1. The **employee\_data.sav** data file on Canvas includes salary and other information about a sample of workers at a mid-size company. The variable salary contains the current salaries of the employees. Perform a one-way analysis of variance to investigate if there are significant differences in salaries among the three job categories. Report your answer below in APA format. Make sure to include all relevant test statistics, the statistical conclusion of the test, post-hoc test results (if relevant) to explore which job categories did or did not differ from one another, and an effect size to quantify the magnitude of the ANOVA effect. Finally, perform and report the Levene’s test to investigate if the data met the equal variance assumption of the ANOVA. Whether the data did or did not meet this assumption, still report the rest of the ANOVA results as described above. [2 points]

* A one-way ANOVA demonstrated a statistically significant main effect of job categories on salaries, F(2, 471) = 434.5, *p* < .001. Posthoc analyses using Tukey’s HSD indicated that salaries for Managers differed significantly from Clerical workers (*p* < .001) and from Custodial workers (*p* < .001), but salaries did not differ significantly between Clerical workers and Custodial workers (*p* = .277). Levene’s test for equality of variances was found to be violated for the present analysis, F(2, 471) = 51.19, *p* < .001.
* **There was a significant difference in salaries among the three job categories, F(2, 471) = 434.5, *p* < .001. The eta-squared was 0.65, indicating a large effect. Tukey’s HSD post-hoc tests demonstrated that managers had significantly higher salaries compared to both clerical (*p* < .001) and custodial workers (*p* < .001). Custodial and clerical workers did not significantly differ (*p =* .28). The Levene’s test for equality of variances was not met, F(2, 471) = 51.19, *p* < .001.**

1. This question uses the same employee data as Q4. Perform a multiple linear regression model with current salary (salary) as the criterion variable and the employee’s beginning salary and time at the job as the two predictors. After running the model, estimate the standardized regression coefficients (betas) and report them below. Next, use the predict() function to estimate what the current salary would be for an employee that had a beginning salary of $50K and has been at the job for 5 years. Provide your answer below. [1 point]

* Employee’s beginning salary and time at the job predicted their current salary, and the multiple regression analysis revealed that the model was significant, R2 = .78, F(2, 471) = 859.4, *p* < .001.
* The standard regression coefficient for an employee’s beginning salary is 0.88, and for the time at the job is 0.10.
* The current salary for an employee that had a beginning salary of $50K and has been at the job for 5 years would be 93907.05.
* **The standardized betas for the beginning salary pre**

1. This question uses the bankloan.sav data file from Q2. Perform a multiple linear regression model with predicted default rate (preddef1) as the criterion. The following four predictors should be included in the model: age, years with current employer, debt to income ratio, and credit card debt. After running the model, assess the assumptions of multicollinearity, normality of the residuals, and equal variance (homoscedasticity) of the residuals. Describe your evaluation of these assumptions below. Include figures to back up your evaluations for the residuals assumptions. Regardless of your evaluation of the assumptions, report the conclusions of the regression model in APA format. This should include a description of the overall significance of the model (e.g., R2, F) and the test statistics and directionality of each of the predictors. [3 points]

* Customer’s age, years with current employer, debt to income ratio, and credit card debt predicted their default rate, and the multiple regression analysis revealed that the model was significant, R2 = .82, F(4, 845) = 975.9, *p* < .001.
* The predictor variables of age (β = -.005, t(845) = -8.39, p < .001), years with the current employer (β = -.02, t(845) = -31.45, p < .001) were negatively related, and debt to income ratio (β = .02, t(845) = 21.87, p < .001) and credit card debt (β = .06, t(845) = 26.60, p < .001) were positively related to the criterion variable predicted default rate.
* Collinearity diagnostics VIF suggests that there is no strong relationship between IVs in the model. Thus, the assumptions of multicollinearity have been met.

Text

Description automatically generated

* The residuals in this model are normally distributed (as shown in the figures below), which suggests that the assumptions of normality of residuals have been met.

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

* The variability of residuals is not equal at all levels of predicted scores, which suggests that the assumptions of equal variance (homoscedasticity) of the residuals have not been met.

Chart, scatter chart

Description automatically generated

1. This last problem uses the **recidivism.sav** data file on Canvas. This data set contains information about individuals that were previously arrested for a crime and whether they have been arrested for a second time (i.e., recidivism). Part of being a good data scientist and statistician is being able to figure out what statistical test is appropriate given a particular research question and the nature of the variables involved. Life will not always be as easy as above where you are told exactly what test to perform. As a result, for this problem, you will have to figure out what statistical test would be appropriate to answer the research question below.

Imagine you are tasked with assessing if there is a relationship between gender and recidivism in this sample. State what statistical test would be appropriate to answer this research question, perform the statistical test, and report your test statistics and conclusion in APA format below. [1 point]

* To assess whether there is a relationship between gender and recidivism, we need to conduct a Chi-square test.
* A chi-square test of independence revealed a non-significant relationship between gender and recidivism in this sample, Χ2(1) = 0.77, *p* = .38.