1. Name and contact information of PI.

Answers will be submitted in Canvas. Not necessary to answer here.

2. Name of the study.

Answers will be submitted in Canvas. Not necessary to answer here.

3. Due date for the analysis.

Answers will be submitted in Canvas. Not necessary to answer here.

4. What is the general purpose of this study/analyses? (exploration, hypothesis confirmation, quality improvement, learning, other) What are the implications for how test-wise and study-wise error rates, inferences from tests, and generalizability of results are handled?

This is an analysis conducted as a learning exercise. The data are fabricated. The results of this analysis could not be generalized to any real population.

5. What are the research questions for this study? Field 9.2.2, 9.9

6. What are the research hypotheses for the research questions?

There were no pre-determined research hypotheses.

7. What is the study design? (Can a specific design described in a study design text like Experimental and Quasi-Experimental Designs for Generalized Causal Inference by Shadish, Cook, and Campbell be identified?) What threats to validity may be applicable to this study design? Are there any procedures that can be put in place to address these threats?

We don’t need to identify this for this learning exercise.

8. What were IRB status and procedures? Any other ethical considerations necessary for this study?

We don’t need to identify this for this learning exercise.

9. When was study data collected?

We don’t need to identify this for this learning exercise.

10. How was the study data collected?

We don’t need to identify this for this learning exercise.

11. What were the data handling procedures?

We don’t need to identify this for this learning exercise.

12. What are the variables in the dataset and the level of measurement for each one?   
Are those set correctly in the dataset?  
A note about the Image variable: I would normally classify that variable as ordinal level of measurement. For purpose of this example, you will need to classify that variable as Continuous. For any other exercise in this class, a variable with 10 or fewer ordered categories should be classified as ordinal level of measurement.

13. Describe data accuracy. Any issues? If so, how will they be addressed? (Some potential things to check about data accuracy might include: data types match levels of measurement, typographical errors in data entry, values which don’t make sense for the phenomena represented by the variable, categories which don’t make sense for the phenomena represented by the variable, decisions about how to address problems identified)

14. Describe any additional data manipulation needed before analysis? (Do any variables need to be reverse coded? Do summary scores for instruments need to be calculated? Are there any other data transformations that need to be done to help meet statistical assumptions?)

15. Describe missing data. Any issues?

16. Describe outliers. Any issues?

17. After initial data cleaning and review, have needed descriptive statistics been calculated before proceeding with analysis? What descriptive statistics will be needed for the study report?

a. Correlations between outcome and predictor variables should be examined before conducting a multiple regression. Create a scatterplot for each of the pairwise combination of the variables.

b. Jamovi – Analysis – Regression – Correlation Matrix

c. Get a plot for the correlation matrix.

18. Will any participant information be included in the study report? (APA Methods Participants section)

We don’t need to identify this for this learning exercise.

19. What statistical test(s) will be conducted? Which variable(s) will be used in those test(s)?

a. The StatKat module can help choose an appropriate test.

b. Jamovi – Analysis – Statkat – Relationships, Prediction, and Group Comparison.

c. Put several continuous variables in the Independent Variables box.

d. Put a continuous variable in the Dependent Variables box.

20. What are the assumptions for the statistical tests employed?

21. Do the data meet those assumptions?

22. NHST steps (for each test):

a. State the null and alternative hypothesis.

b. Establish the criteria for rejection (alpha level).

c. Calculate the test statistic.

d. Draw conclusion about the null.

e. Conduct post-hoc analyses if any.

f. Report results.

a. What is the null and alternative hypotheses for multiple regression?

i. Hint: every time a p-value is calculated there are associated null and alternative hypotheses.

ii. What are the null and alternative hypotheses for the F test(s) of the model?

iii. What are the null and alternative hypotheses for the t-test(s) for the model parameters? (intercept, b for each predictor variable)

b. What is the criteria for rejection?

i. There are situations where you may want to use a different alpha level than alpha = .05.

ii. If you're doing pilot studies you're mostly looking for trends and possibilities. The alpha level may be relaxed to alpha = .10.

iii. If you're doing something like drug studies where there could be negative consequences for false positives the alpha level may be more strict and lowered to alpha = .01.

iv. If you’re doing multiple statistical tests on a data set, different strategies to correct for study-wise error rate may be used to lower the alpha level for individual tests. (like a Bonferroni correction)

v. We're going to stick to alpha = .05 for class exercises.

c. Fit the multiple regression model using Adverts, Airplay, and Image to predict Sales .

i. Jamovi - Analyses - Regression - Linear Regression.

ii. Move Sales to the Dependent Variable box

iii. Move Adverts, Airplay, Image to the Covariates box

iv. Model Builder: Open the Model Builder section. Click on Add New Block. Drag Airplay and Image from Block 1 to Block 2. (We’re going to do a hierarchical regression like Field in section 9.10.)

v. Assumption Checks: Check everything except Shapiro-Wilk.

vi. Model Fit: Check everything.

vii. Model Coefficients: Check everything.

d. What is your decision about the null hypotheses for the multiple regression?

e. Use regression diagnostics to assess the model.

i. Bias in the model

1. Outliers

a. Residuals

b. Standardized residuals

i. Jamovi (could look at Q-Q Plot of residuals)

c. Studentized residuals

2. Influential cases

a. Cook’s distance (overall influence of a case in the model)

i. Jamovi in Data Summary

b. Leverage or hat values (influence of observed value over predicted)

i. not in Jamovi

c. Mahalanobis distance (distance of cases from means of predictors)

i. not in Jamovi

d. Standardized dfBeta (changes in b from excluding a case)

i. not in Jamovi

e. Standardized DFFit (difference in predicted value when including/excluding case)

i. not in Jamovi

f. CVR (covariance ratio – case influences variance or parameters)

i. not in Jamovi

ii. Generalizability of the model

1. Additivity and linearity

a. Residual plots for each predictor

2. Independence of errors

a. Durbin-Watson test

3. Homoscedasticity

a. Residual plots for each predictor

4. Normally distributed errors

a. Q-Q Plot

5. Predictors uncorrelated with external variables

6. Correct variable types

a. Known from study design

7. Multicollinearity

a. VIF, Tolerance

b. Check correlations between predictors < .8

8. Non-zero variance

a. Check descriptive statistics for predictors

9. Cross validation

iii. Adequate sample size

1. Known from study design, effect size, # predictors assessment

iv. What are the purposes of regression diagnostics? Field 9.3 – 9.4

f. Read Field chapter 9.11 for a nice explanation of how to interpret the results for this multiple regression. Make particular note of how he interprets regression diagnostics in section 9.11.5 – 9.11.7. The output for many types of regression analyses will have these same basic parts (overall model fit & model parameters). The particular tests reported in these two sections may differ depending on the type of regression analysis, but the basic information is very similar.

One of the areas where Jamovi development has not quite achieved a mature level is in regression diagnostics. Run the code in the R notebook .rmd file to see how a multiple regression model can be fit in R and some available regression diagnostic tests and plots.

i. Getting a number of regression diagnostics quickly

1. broom::augment(model)

ii. Bias in the model

1. Outliers

a. Residuals

i. resid(model)

b. Standardized residuals

i. rstandard(model)

c. Studentized residuals (differ from other data)

i. MASS::studres(model)

ii. rstudent()

iii. look for number of cases greater than 2 or 3

d. plots for outliers

i. car::outlierTest(model)

ii. qqPlot(model)

iii. leveragePlots(model)

e. standardized residuals with lines at cutoff points

i. res.std <- rstandard(model)

ii. plot(res.std, ylab = "Standardized Residual", ylim=c(-3.5, 3.5))

iii. abline(h = c(-3, 0, 3), lty=2)

iv. index <- which(res.std > 3 | res.std < -3)

v. text(index-20, res.std[index], labels = dataset$id[index])

vi. print(index)

2. Influential cases

a. Cook’s distance (overall influence of a case in the model)

i. cooks.distance(model)

ii. looks for values greater than 1

iii. plot(model, which = 4, cook.levels = cutoff)

iv. influencePlot(model, main="Influence Plot", sub="Circle size proportional to Cook's Distance")

v. infIndexPlot(model)

b. Leverage or hat values (influence of observed value over predicted)

i. hatvalues(model)

ii. look for values greater than 2 to 3 times the average

iii. plot for high leverage values

1. #a vector containing the diagonal of the 'hat' matrix

2. h <- influence(model)$hat

3. #half normal plot of leverage from package faraway

4. halfnorm(influence(model)$hat, ylab = "leverage")

c. Mahalanobis distance (distance of cases from means of predictors)

i. mahalanobis(dataset[,c(“pred1”,”pred2”)],colMeans(dataset[,c("Adverts","Airplay","Image")]),cov(dataset[,c("Adverts","Airplay","Image")]))

ii. look for values greater than cutoff for chi-square with df = number of predictors

d. Standardized dfBeta (changes in b from excluding a case)

i. dfbetas(model)

ii. look for standardized values greater than 1

e. Standardized DFFit (difference in predicted value when including/excluding case)

i. dffits(model)

ii. look for standardized values greater than 1

f. CVR (covariance ratio – case influences variance or parameters)

i. covratio(model)

ii. look for values outside of bounds (1 +/- (k\*(K+1)/n))

g. function to help flag influential cases

i. stats::influence.measures(model)

h. plots for finding influential cases

i. avPlots(model,”predictorName”)

ii. crPlots(model,”predictorName”)

iii. Generalizability of the model

1. Additivity and linearity

a. First plot in plot(model)

b. car::residualPlots(model)

2. Independence of errors

a. Plot residuals with ID plot(model$resid ~ dataset$id)

b. durbinWatsonTest(model)

c. dwt(model)

3. Homoscedasticity

a. First plot in plot(model)

b. Breusch-Pagan test in bptest(model)

4. Normally distributed errors

a. Second plot in plot(model)

b. hist(dataset$m3studres)

c. shapiro.test(resid(model))

5. Predictors uncorrelated with external variables

6. Correct variable types

7. Multicollinearity

a. vif(model.3)

b. 1/vif(model.3)

c. mean(vif(model.3))

8. Non-zero variance

9. Cross validation

iv. Adequate sample size

v. What are the purposes of regression diagnostics? Field 9.3 – 9.4

vi. Additional resources

1. Field authored a version of his text using R a number of years ago. The section on regression lists some of the R functions which are useful for regression diagnostics. Chapter 7.9.2 https://archive.org/details/discoveringstatisticsusingr/page/n321/mode/2up

2. Applied Statistics with R Ch. 13 (optional) available at: https://daviddalpiaz.github.io/appliedstats/model-diagnostics.html

3. R for Researchers OLS (optional) available at: https://ssc.wisc.edu/sscc/pubs/RFR/RFR\_Regression.html

4. R for Researchers regression diagnostics (optional) available at: https://ssc.wisc.edu/sscc/pubs/RFR/RFR\_Diagnostics.html

5. Diagnostics in multiple linear regression (optional) available at: https://web.stanford.edu/class/stats191/notebooks/Diagnostics\_for\_multiple\_regression.html

6. Introduction to regression diagnostics (optional) available at: https://stats.idre.ucla.edu/wp-content/uploads/2019/02/R\_reg\_part2.html#(1)

7. Regression model diagnostics (optional) available at: http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

8. Regression diagnostics (optional) available at: https://www.statmethods.net/stats/rdiagnostics.html

9. Outlier detection with Mahalanobis distance (optional) available at: https://www.r-bloggers.com/outlier-detection-with-mahalanobis-distance/

10. Multivariate outlier (optional) available at: https://en.wikiversity.org/wiki/Multivariate\_outlier