Question 1 (8 parts): Several years ago, a Chinese automobile company aspired to enter the US market by developing manufacturing capabilities and producing cars locally. They contracted with a major automobile consulting company to better understand how the drive train affected the pricing of cars in the American market, which they suspected would be different from the Chinese market.

The Workbook contains a sample of 185 foreign and domestic automobiles that were being sold in the US. The variables are: price (in US dollars) engine size (in cc [cubic centimeters]) horsepower (in bhp [brake horsepower]) cylinders turbo (1=yes; 0=no) mpg (average miles per gallon) The variables are in the order listed above.

The sample data is provided in two Question 1 tabs in the Workbook.

A. Using the first tab, labeled “Question 1 - parts a,b,c”, build a model that relates price to the other variables listed above. Write the estimated regression equation.

B. Is the model statistically significant at the α = .05 level? Write out the null and alternative hypotheses, the appropriate test statistic, the p-value for that test statistic and your conclusion, based on that p-value.

C. The automobile company suspects that vehicle price depends on engine size, holding everything else equal. To evaluate this suspicion, write out the null and alternative hypotheses, the appropriate test statistic, the p-value for that test statistic and your conclusion based on that p-value, at the α = .05 level.

D. The automobile company also suspects that prices are different for turbocharged cars, holding everything else equal. Interpret the slope of turbo to determine whether the sample data supports this suspicion. Specifically, report the appropriate test statistic, the p-value for that test statistic and your conclusion, based on that p-value, at the α = .05 level.

E. Another component of the design factor which was thought to affect pricing was the drive wheels—front wheel drive (fwd), rear wheel drive (rwd), or four wheel drive (4wd). In the second tab, labeled “Question 1 - parts d,e,f,g”, you will find the variable drive wheels. Setting front wheel drive (fwd) as the baseline, construct dummy variables to capture drive wheels, then add these dummy variables to the model you used in the previous regression and run a regression for this new model. Write the estimated regression equation of this model.

F. In the model you built in part (e), is the price different for vehicles with rear wheel drive (rwd) compared to front wheel drive (fwd) at the α = .05 level, holding everything else equal? Cite the appropriate test statistic and p-value from your Excel output. g. Now select a model to predict vehicle price by applying the backward elimination approach. What is your selected model’s adjusted R 2 ? List the independent variables included in this model.

H. The company considered pricing its best-selling model at $15,000. That model used front wheel drive (fwd) and had a 125 cc, 6 cylinder, non-turbocharged engine making 100 horsepower and 18 mpg. Use the regression equation from (g) above to compute the vehicle’s expected price. Then develop an approximate 80% prediction interval for the vehicle’s price. Does the proposed $15,000 price fall outside that prediction interval?

Question 2 (4 parts): On November 20, 2020, Pfizer and BioNTech applied to the US Food and Drug Administration for emergency use authorization of their Covid-19 vaccine.

1 The application was based on interim results from the phase 3 clinical trial. Of the 170 participants in the trial who became infected with Covid-19, 8 had received the vaccine while 162 had not (they had received a placebo instead of the vaccine). Note that study participants who were infected with Covid were exposed naturally just like everyone else in the population over the period of the clinical trial. Assume that 21,000 of the 41,135 study participants had received the vaccine while the remainder had not.

A. In the Workbook, you will find a 2×2 contingency table with vaccine received (Yes/No) by row and subsequent Covid infection (Yes/No) by column. Complete the contingency table with joint and marginal frequencies (not probabilities) from the information above.

B. Compute the expected frequencies, assuming independence, of receiving the vaccine (Yes/No) and subsequent Covid infection (Yes/No).

C. Use the expected and actual frequencies to determine if Covid infection depends on whether or not one had received the vaccine. Specifically, report the appropriate test statistic, the p-value for that test statistic and your conclusion, based on that p-value.

D. First, compute the probability that a participant who had received the vaccine was subsequently infected with Covid. Second, compute the probability that a participant who was subsequently infected with Covid had received the vaccine. Which of these probabilities, the first or the second, offers information about the effectiveness of the vaccine compared to the placebo (i.e., no vaccine)?

Question 3 (5 parts): Consider two different stocks, X and Y, with expected returns μX = 17.0% and μY = 14.9% and with standard deviations σX = 12.4% and σY = 10.0% for those returns. The two assets have a correlation, ρ, of -0.7.

A. What is the covariance of the returns of stocks X and Y? (Hint: use the definition of the population correlation ρ to compute the covariance) In general, will constructing a portfolio from these two stocks reduce or increase the risk compared to the individual stocks? Briefly explain.

B. What is the expected return and standard deviation of a portfolio made up of stocks X and Y which is 20% stock X (the remainder stock Y)?

C. What is the expected return and standard deviation of a portfolio made up of stocks X and Y which is 50% stock X?

D. What is the expected return and standard deviation of a portfolio made up of stocks X and Y which is 80% stock X?

E. Which of the portfolios above, either (b), (c), or (d), offers the best combination of risk and return? Briefly explain.

Question 4 (8 parts): A small online retailer of home décor and art compiled time series data that allows us to predict net dollar sales. Three years of monthly data is provided in the Question 4 tab of the Workbook, including the following variables: month number of orders returns (in dollars) discounts (in dollars) net sales (in dollars)

A. Determine the sample correlation coefficient, r, between number of orders and net sales. Test the alternative hypothesis that number of orders has a linear relationship to net sales. Specifically, what are the test statistic and the p-value for that test statistic? For α = .05, what do you conclude about the relationship between the variables?

B. Now estimate a simple linear regression to predict the net sales using only discounts. Write the estimated regression equation.

C. How well does the model fit the dependent variable? Cite the appropriate statistic from the regression output, and interpret that statistic.

D. What is the fitted slope coefficient? Interpret that slope coefficient. Is there a linear relationship between the independent and dependent variables at α = .05? Report the relevant test statistic and p-value for that test statistic.

E. Now estimate a multiple linear regression to predict net sales using number of orders, discounts, and returns. Write the estimated regression equation.

F. Is the model estimated in part (e) statistically significant at the α = .05 level? Write the null and alternative hypotheses for this test, the appropriate test statistic, the p-value for that test statistic, as well as your conclusion.

G. Using the model estimated in part (e), which independent variables (if any) are related to the dependent variable net sales at the α = .05 level, in this model? Include relevant values from the Excel output to support your conclusions.

H. Compare the estimated slope coefficients for the variable discounts in the simple linear regression in parts (b) to (e) and the multiple linear regression in parts (e) to (g). Are they different? Briefly explain why or why not.