Question 1 *(20 points)*

A number of retailers regularly survey their customers to determine among other things, whether they were happy with their purchase or service and whether they intended to return. A chain of hardware stores/automobile service centers is one such company. At the completion of repair work customers are asked to fill out the following form (shown in figure 1):

A random sample of 134 responses was drawn. The responses to questions 1 through 4 (1 = poor, 2 = fair, 3 = good, 4 = very good) are stored in columns 1 through 4, respectively. Responses to question 5 (2 = yes, 1 = no) are stored in column 5. Column 6 stores a 1 if a positive comment was made, 2 if a negative comment was made, and 3 if no comment was made. The data set is available on the CSV file *Automobile Service Center Ratings*.

	Tell us what you think.				
	Are You Satisfied?	Very Good	Good	Fair	Poor
1	Quality of work performed				
2	Fairness of price				
3	Explanation of work and guarantee				
4	Checkout process				
5	Will return in future YES NO				
	Comments?				

Figure 1: Question 5 Survey

a) *10 points* Can we infer that those who say will return assess each category higher than those who will not return? The meaning of this question is: For each of the four questions on the survey, can we infer that those who say will return rate each category (question) higher than those who will not return?

b) *10 points* Is there sufficient evidence to infer that those who make positive comments, negative comments, and no comments differ in their assessments of each category?

Question 2 *(10 points)*

A well-known soft-drink manufacturer has used the same secret recipe for its product since its introduction over 100 years ago. In response to a decreasing market share, however, the president of the company is contemplating changing the recipe. He has developed two alternative recipes. In a preliminary study, he asked 20 people to taste the original recipe and the two new recipes. He asked each to evaluate the taste of the product on a 5-point scale, where 1 = Awful, 2=Poor,3=Fair,4=Good, and 5=Wonderful. The dataset is available on the CSV file *Soft Drink Recipe*. The president decides that unless significant

differences exist between evaluations of the products, he will not make any changes. At 5% significance level, use an appropriate statistical test to conclude if there are any differences in the ratings of the three recipes.

Question 3 *(10 points)*

Refer to the dataset from the General Social Survey provided on the CSV file *Job Loss*. If one works longer hours (`HRS1`) does the chances of losing one's job (`JOBLOSE: 1 = Very likely, 2 = Fairly likely, 3 = Not too likely, 4 = Not likely`) become less likely? Conduct an appropriate statistical test to answer the question.

Question 4 *(10 points)*

Does the brand name of an ice cream affect consumers' perceptions of it? The marketing manager of a major dairy pondered this question. She decided to ask 60 randomly selected people to taste the same flavor of ice cream in two different dishes. The dishes contained exactly the same ice cream but were labeled differently. One was given a name that suggested that its maker was European and sophisticated; the other was given a name that implied that the product was domestic and inexpensive. The tasters were asked to rate each ice cream on a 5-point scale, where 1 = Poor, 2 = Fair, 3 = Good, 4 = Very good, and 5 = Excellent. Do the results allow the manager to conclude at the 10% significance level that the European brand is preferred? The data set is available on the CSV file *Ice Cream Comparison*.

Question 5 *(10 points)*

A locksmith is in the process of selecting a new key-cutting machine. If there is a difference in key-cutting speed between the two machines under consideration, he will purchase the faster one. If there is no difference, he will purchase the cheaper machine. The times (in seconds) required to cut each of the 24 most common types of keys were recorded. The data set is available on the CSV file *Machine Selection*. What should he do?

Question 6 *(10 points)*

It is often useful for companies to know who their customers are and how they became customers. In a study of credit card use, random samples were drawn of cardholders who applied for the credit card and credit cardholders who were contacted by telemarketers or by mail. The total purchases made by each last month were recorded. Can we conclude at the 5% significance level from these data that differences exist on average between the two types of customers? Navigate the flowchart and choose the appropriate statistical test to run. The data set is available on the CSV file *CreditcardHolders*.

Question 7 *(10 points)*

Are Americans more deeply in debt this year compared to last year? To help answer this question a statistics practitioner randomly sampled Americans this year and last year. The sampling was conducted so that the samples were matched by the age of the head of the household. For each, the ratio of debt payments to household income was recorded. Can we infer at the 5% significance level that the ratios are higher this year than last? Navigate the flowchart and choose the appropriate statistical test to run. The data set is available on the CSV file *AmericanDebt*.

Question 8 *(10 points)*

High turnover of employees is expensive for firms. The firm not only loses experienced employees, it must also hire and train replacements. A firm is considering several ways to improve its retention (the proportion of employees who continue with the firm after 2 years). The currently favored approach is to offer more vacation days. Improved health benefits are a second alternative, but the high cost of health benefits implies that to be effective this benefit must increase retention by 0.05 above that associated with offering increased vacation days. To choose between these, a sample of 125 employees on the West Coast was given increased health benefits, and a sample of 140 on the East Coast was offered increased vacation time. Data set is available on the file *Benefits Comparison*.

a) What are potential confounding effects in this comparison?

b) Do the data indicate that offering health benefits has statistically significantly higher retention to compensate for switching to health benefits?

c) Is there a statistically significant difference in retention rates between the benefit plans?

Question 9 *(10 points)*

Use the *Wage* data set from the class solved examples (also included in the Homework 4 folder) and apply the Spearman's rank correlation test for heteroscedasticity to find out of there is evidence of heteroscedasticity in the data.

Question 10 *(10 points)*

For the data given in the CSV file *Compensation*, regress average compensation Y on average productivity X, treating employment size as the unit of observation.

a) From the preceding regression obtain the residuals $\hat{u}_i$.

b) Following the Park test, regress ln $\hat{u}_i{}^2$ on ln X_i and verify if there is heteroscedasticity problem.

c) Following the Glejser approach, regress $|\hat{u}_i|$ on X_i and then regress $|\hat{u}_i|$ on $\sqrt{X_i}$. Verify if there is heteroscedasticity problem.

d) Use the Spearman rank correlation test to verify if there is a heteroscedasticity problem.

Question 11 *(12 points)*

The data set on the CSV file *R&D* gives data on research and development (R&D) expenditure, sales, and profits for 14 industry groupings in the United States (all figures in millions of dollars). Since the cross-sectional data presented in this table are quite heterogeneous, in a regression of R&D on sales, heteroscedasticity is likely. Use the Park test, Glejser test, and White's test to assess if there is heteroscedasticity problem.

Question 12 *(10 points)*

Techcore is a high-tech company located in Forth Worth, Texas. The company produces a part called fiber-optic connector (FOC) and wants to generate reasonable accurate but simple forecasts of sales of FOCs over time. The company has weekly sales data for the past 265 weeks on the CSV file named *FOC*. The data have been disguised to provide confidentiality. Verify using informal and formal tests that heteroscedasticity is a problem. Transform the $SALES$ variable to $\ln(SALES)$ and check to see if the variances have been stabilized. Using the transformed model forecast the sales in week 300.

Question 13 *(10 points)*

One of the very common detection techniques for heteroscedasticity is **Breusch-Pagan** test. This is a method of testing for heteroscedasticity in the error term by investigating whether the squared residuals can be explained by possible proportionality factors. (I have posted a paper "A Simple Test for Heteroscedasticity and Random Coefficient Variation" by Breusch and Pagan which was published in *Econometrica*, Vol 47, pp. 1287-1294. This paper is available on Canvas for those who are interested in reading about the theory.) Here is how the procedure works. Say your original regression model contains two regressors as shown below.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon$$

1. Obtain the residuals from the estimated regression equation.

2. Use the squared residuals as the regressand in an auxiliary regression which contains all the independent variables as regressors.

$$u_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \nu_i$$

3. Test the overall significance of the equation in step 2 above with a chi-square test. The null and alternative hypotheses are:

$$H_0 : \alpha_1 = \alpha_2 = 0$$

$$H_1 \text{ At least one } \alpha \text{ not equal to zero}$$

The null hypothesis is homoscedasticity, because if α_1 and α_2 are both zero, then the variance equals α_0 which is a constant. The test statistic is N times R^2, or the sample size (N) times the unadjusted R^2 from the equation in step 2. The test statistic has a chi-square distribution with degrees of freedom equal to the number of slope coefficients in the auxialiary regression (in step 2). If NR^2 is greater than or equal to the critical chi-square value (or the p-value is less than any chosen α), then we reject the null hypothesis of homoscedasticity.

Suppose that you have been hired as an analyst to determine the best location for the next Woody's restaurant, where Woody's is a moderately priced family restaurant chain. (The data in this example are real. They are from a sample of 33 Denny's restaurants in Southern California), but the number of independent variables considered is much smaller than was used in the actual research.) CSV Data file *Woody*. You decide to build a regression model to explain the gross sales volume at each of the restaurants in the chain as a function of various descriptors of the location of that branch. If you can come up with a sound equation to explain gross sales as a function of location, then you can use this equation to help Woody's decide where to build their newest eatery. The variables used in the exercise are:

- Y = the number of customers served (measured by the number of checks or bills that the servers handed out) in a given location in the most recent year.
- N = Competition: the number of direct market competitors within a two-mile radius of the Woody's location
- P = Population: the number of people living within a three-mile radius of the Woody's location
- I = Income: the average household income of the population measured in variable P

a) Check using the Breusch-Pagan test if the condition of homoscedasticity is satisfied.

b) The `lmtest` package in `R` includes a function called `bptest`. Verify your results obtained using the three step approach with that obtained using the `bptest` function in `R`.

c) Run the White's general heteroscedasticity test to validate your findings from the parts a and b above. Remember to include all the square and cross-product terms.

d) Does the Koenker-Bassett heteroscedasticity test lead to findings consistent with parts a), b), and c) above?

Question 14 *(12 points)*

The data obtained from a survey of 9,966 economists in 1964 are given on the CSV file *EconomistSalary*.

a) Develop a suitable regression model to explain median salary in relation to age. For the purpose of regression, assume that median salaries refer to the midpoint of the age interval.

b) Assuming error variance proportional to age, transform the data and obtain the WLS regression.

c) Now assume that it is proportional to the square of age. Obtain the WLS regression on this assumption.

d) Plot the residuals from regressions b) and c) and see if they exhibit any systematic patterns. If they do, use the Park or Glejser test to further confirm if there is evidence of heteroscedasticity in the data.

Question 15 *(10 points)*

Christmas week is a critical period for most ski resorts. Because many students and adults are free from other obligations, they are able to spend several days indulging in their favorite pastime, skiing. A large proportion of gross revenue is earned during this period. A ski resort in Vermont wanted to determine the effect that weather had on its sales of lift tickets. The manager of the resort collected data on the number of lift tickets sold during Christmas week (Y), the total snowfall in inches (X_1), and the average temperature in degrees Fahrenheit (X_2) for the past 20 years. Develop the multiple regression model and diagnose any violation of independence, normality, and constant variance. Use formal and informal tests for all conditions; i.e., formal test and informal test for normality, formal test and informal test for independence, and formal and informal test for constant variance. In your model, include an independent variable that has a time-ordered effect on the dependent variable. To this end, include a new variable *Time Periods* as an additional explanatory variable. This variable is a list of sequential time periods. Run the multiple regression with this new variable included. Reassess the required conditions and see if any violations noticed in the previous model got remedied by including the time-ordered effect. Data available on the CSV file *SkiSales*.

Question 16 *(15 points)*

Consider the following model of wage determination:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$$

where Y = wages = index of real compensation per hour and X=productivity=index of output per hour.

a) Using the data on the CSV file *CompensationAndProductivity*, find the regression of Y on X. Is there evidence of autocorrelation? Create a residual by time plot and the Durbin-Watson statistic to answer this question. Create and include the time index along with X. Does including the time index change anything related to autocorrelation?

b) Using the data on the CSV file *CompensationAndProductivity*, estimate the above model ($Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$) and interpret your results.

c) Since the model contains lagged regressand as a regressor, the Durbin–Watson d is not appropriate to find out if there is serial correlation in the data. For such models, called **autoregressive models**, Durbin has developed the so-called ***h* statistic** to test for first-order autocorrelation, which is defined as:

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n[var(\hat{\beta}_3)]}}$$

where n=sample size, $var(\hat{\beta}_3)$=variance of the coefficient of the lagged Y_{t-1}, and $\hat{\rho}$ = estimate of the first-order autocorrelation.

For large sample size (technically, asymptotic), Durbin has shown that, under the null hypothesis that $\rho = 0$, h $N(0, 1)$, that is, the h statistic follows the standard normal distribution. From the properties of the normal distribution we know that the probability of $|h| > 1.96$ is about 5 percent. Therefore, if in an application $|h| > 1.96$, we can reject the null hypothesis that $\rho = 0$, that is, there is evidence of first-order autocorrelation in the autoregressive model given above.

To apply the test, we proceed as follows: *First*, estimate the above model by OLS (don't worry about any estimation problems at this stage). *Second*, note $var(\hat{\beta}_3)$ in this model as well as the routinely computed d statistic. *Third*, using the d value, obtain $\hat{\rho} = (1 - d/2)$. It is interesting to note that although we cannot use the d value to test for serial correlation in this model, we can use it to obtain an estimate of ρ. *Fourth*, now compute the h statistic. *Fifth*, if the sample size is reasonably large and if the computed $|h| > 1.96$ exceeds 1.96, we can conclude that there is evidence of first-order autocorrelation. Of course, you can use any level of significance you want.

Apply the h test to the autoregressive wage determination model given earlier and draw appropriate conclusions and compare your results with those given in regression in part a above.

Question 17 *(15 points)*

Does dieting affect the brain? if so, how? This question was addressed by researchers. The experiment used 40 middle-age women in Adelaide, Australia; half were on a diet and half were not. The mental arithmetic part of the experiment required the participants to add two three-digit numbers. The amount of time taken to solve the 48 problems was recorded. The participants were given another test that required them to repeat a string of five letters they had been told 10 seconds earlier. They were asked to repeat the test with five words told to them 10 seconds earlier. The data were recorded in the following way:

Column A: Identification number
Column B: 1 = dieting, 2 = not dieting
Column C: Time to solve 48 problems (seconds)
Column D: Repeat string of 5 letters (1 = no, 2 = yes)
Column E: Repeat string of 5 words (1 = no, 2 = yes)

At 0.05 significance level, is there sufficient evidence to infer that dieting adversely affects the brain? Run appropriate tests to answer the question. The data set is on the CSV file *DietEffect*.

Question 18 *(15 points)*

John Maynard Keynes, one of the most influential economists since Adam Smith, developed the notion of a consumption function, which explains total consumption as a function of disposable personal income. Our goal is to model U.S. aggregate consumption as a function of disposable personal income and the real interest rate. The data are from the St. Louis Federal Reserve FRED database and the *Economic Report of the President*. Descriptions of the variables are given below, along with the hypothesized signs for the coefficients, and the data set itself is on the the CSV file *Consumption*.

Variable	**Description**	**Expected sign**
con_t	Real personal consumption expenditures in year t, in billions of 2009 dollars	NA
dpi_t	Real disposable personal income in year t, in billions of 2009 dollars	+
aaa_t	The real interest rate on AAA corporate bonds in year t	-
$year_t$	Year t	NA

Before assigning, I worked this entire question in `R` and IT WORKS!

a) Estimate the consumption function, using disposable personal income and the real interest rate as the independent variables.

b) Generate the residuals from the regression part a) (naming them $\hat{u}_i$) and plot them as a line graph against time period t (with t on the X-axis). Does the plot look entirely random? For this part, you will need to create the time period vector in the data set, which is just a sequence of numbers from 1 to 62.

c) Conduct a Durbin–Watson test for positive serial correlation at the 5% significance level. Carefully write down the null and alternative hypotheses.

d) Let's see if our Durbin–Watson results can be confirmed with the Lagrange Multiplier test (Breusch-Godfrey test). Conduct a Lagrange Multiplier test (Breusch-Godfrey test) for autocorrelation at the 5-percent level. What can you conclude? Use an $AR(1)$ scheme.

e) If you encountered first-order positive autocorrelation in either of the previous steps, re-estimate our aggregate consumption model using Generalized Least Squares. Use the Prais-Winsten transformation for the first observation. Are the GLS coefficients and t-statistics the same as the OLS coefficients and t-statistics? Why?

f) After the GLS transformation, does autocorrelation still appear to exist? Support your answer using the Durbin-Watson test.

g) Now estimate the aggregate consumption model using the Newey-West method with a lag of 1. For the Newey-West method, use the `sandwich` package in `R`. Here is the reference. https://bookdown.org/machar1991/ITER/15-4-hac-standard-errors.html The `NeweyWest()` function in `R` will display the variance-covariance matrix. You want to see the adjusted standard errors of the coefficients. While the site shows you how to calculate the standard errors from the matrix, but using the `coeftest` function in the `lmtest` package is much convenient. This will display the coefficients of all the parameter estimates + the "adjusted" standard errors.

h) After the Newey-West calculation, are the coefficients the same as the OLS coefficients? What is the difference between the Newey-West output and the OLS output?