Question 1 *(20 points)*

Suppose that Electronics World, a chain of stores that sells audio and video equipment, has gathered the data in CSV file *Electronics*.

These data concern

- store sales volume in July of last year (Y, measured in thousands of dollars)
- the number of households in the store's area (X, measured in thousands)
- the location of the store (on a suburban street, in a suburban shopping mall, or in downtown — a qualitative independent variable)

Create a multiple regression model with Sales as the response variable, number of households and Location as the regressors. Set the `Street` location as the base or the reference category. Compare the sales of the three locations using the point estimates and the 95% confidence interval estimates. This question is a practice on interpreting dummy variable coefficients. I don't require drawing any inferences for this question. You will not be penalized for not drawing any inferences. In other words, no Global F-test and no individual t-tests for beta coefficients are necessary.

Question 2 *(30 points)*

The *WinterFun* Company sells winter sports merchandise including skis, ice skates, sleds, and so on. Quarterly sales (in thousands of dollars) for the *WinterFun* company are shown on the CSV file *WinterFun*. The time period represented starts in the first quarter of 2008 and ends in the fourth quarter of 2017.

A linear regression containing only the time variable like shown below is called a **linear trend model**.

$$y = \beta_1 + \beta_2 t + \varepsilon$$

One of the common objectives of such analysis is if the linear trend model is sufficient for predicting sales or are the sales influenced by seasonality. For this question, follow the steps provided in the parts below to construct a final model:

a) Create a linear trend model. Write the model. Linear trend model means using only time (t) as the regressor.

b) Now, you want to see if there is any seasonality. For this part, let's use a descriptive approach. Create a time plot of sales (Sales versus Time) and check if there are seasonal patterns. Examine evidence of seasonality visually. Describe the seasonality in your own words.

c) Create indicator variables for quarters.

d) Conduct a Partial F-test to assess if the seasonal indicator variables are necessary in the model. Write the full model, the reduced models, and the hypotheses clearly. The reduced model in this case is the linear trend

model and the full model is the one with indicator and the trend variables. Write your conclusion and interpretation. Compute the coefficient of partial determination and interpret it.

Question 3 *(30 points)*

The dataset on *EmploymentDiscrimination* file presents data from the case of *United States Department of the Treasury v. Harris Trust and Savings Bank* (1981). The data includes the salary of 93 employees of the bank (SALARY), their education level (EDUCAT), and their gender (GENDER).

a) Create a multiple regression model using Salary as the regressand and education level and gender as the regressors. For consistency, let's use the `Male` dummy, although as you know, which dummy you use won't matter.

b) Interpret the differential intercept coefficient and the parameter estimate of the education level. Is there evidence of employment discrimination at the Harris bank?

c) Does the difference in average salaries increase between two groups as education increases? Note: this question means that you want to test for the interaction between education and gender. An alternate formulation of the question is: Does the effect of gender on salary depend on the level of education? Create a new model to answer this question.

d) Create a plot with the two regressions – one for male and another for female. Are these two regressions parallel, coincident, dissimilar, or concurrent?

e) What is the difference between models in part a and part c? Your comparison should include examining the statistical significance of the variables, the adjusted R^2, the model standard error, the overall model validity, and the t-test for individual coefficients.

f) Now run a partial F-test to assess the significance of the gender dummy and the interaction term. Are the gender dummy and the interaction term jointly significant in explaining the variation in salaries?

g) Explain what causes the conflicting results in parts e and f above. In other words, why is the statistical significance conflicting between the individual t-test and the Partial F-test? Which model would you settle with?

Question 4 *(10 points)*

Before purchasing videoconferencing equipment, a company ran tests of its current internal computer network. The goal of the tests was to measure how rapidly data moved through the network given the current demand on the network. Eighty files ranging in size from 20 to 100 megabytes (MB) were transmitted over the network at randomly chosen times of day, and the time to send the files (in seconds) recorded. Two types of software were used to transfer the files, identified by the column labeled Vendor in the data table. The two

possible values are "MS" and "NP". Compare the download times produced by the two vendors using an analysis of covariance that takes account of the differences in file sizes. Summarize the comparison of download times based on this analysis. Also, does the effect of file size on transfer time different for each vendor? Which vendor would you recommend? Dataset available under CSV file *Downloads*.

Question 5 *(10 points)*

The data set in the CSV file *Fisher Index* gives data on the annual rate of return Y (%) on a mutual fund and a return on a market portfolio as represented by the Fisher Index, X (%). Now consider the following model, which is known in the finance literature as the *characteristic line*.

$$Y_t = \beta_1 + \beta_2 X_i + u_i$$

In the literature there is no consensus about the prior value of β_1. Some studies have shown it to be positive and statistically significant and some have shown it to be statistically insignificant. In the latter case, the above model becomes a regression-through-the-origin model, which can be written as

$$Y_t = \beta_2 X_i + u_i$$

Use the data given to estimate both these models and decide which model fits the data better.

As discussed in the class briefly, for the regression-through-the-origin regression model the conventionally computed R^2 may not be meaningful. One suggested alternative for such models is the so-called "raw" R^2, which is defined (for the two-variable case) as follows:

$$Raw\ R^2 = \frac{(\Sigma X_i Y_i)^2}{\Sigma X_i^2 \Sigma Y_i^2}$$

If you compare the raw R^2 with the conventional R^2, you will see that the sums of squares and cross-products in the raw R^2 are not mean-corrected. For the interceptless model shown above, compute the raw R^2. Compare this with the R^2 value that you obtained from the model with intercept. What conclusions can you draw?

Question 6 *(10 points)*

The data set on the CSV file *CorporateFinancials* gives data on after-tax corporate profits and net corporate dividend payments ($, in billions) for the United States for the quarterly period of 1997:1 to 2008:2.

a) Regress dividend payments (Y) on after-tax corporate profits (X) to find out if there is a relationship between the two.

b) To see if the dividend payments exhibit any seasonal pattern, develop a suitable dummy variable regression model and estimate it. In developing the model, how would you take into account that the intercept as well as the slope coefficient may vary from quarter to quarter?

c) Based on your results, what can you say about the seasonal pattern, if any, in the dividend payment policies of U.S. private corporations? Is this what you expected a priori?

Question 7 *(10 points)*

The marketing manager at *CleanLawns*, a lawn mower company, believes that monthly sales across all outlets (stores, online, etc.) are influenced by three key variables: (1) outdoor temperature (in °F), (2) advertising expenditures (in $1,000s), and (3) promotional discounts (in %). The CSV data file *Mowers* shows monthly sales data over the past two years.

a) Estimate the model

$$Sales = \beta_0 + \beta_1\ Temperature + \beta_2\ Advertising + \beta_3\ Discount + u.$$

Test for the joint and individual significance of the explanatory variables at the 5% level.

b) Examine the data for evidence of multicollinearity. Provide two reasons why it might be best to do nothing about multicollinearity in this application.