1. [Conceptual] For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small. We would generally expect the performance of a flexible statistical learning method to be better than an inflexible method because a flexible method is going have a better fit for a larger sample size/greater number of observations than will an inflexible method.

(b) The number of predictors p is extremely large, and the number of observations n is small. We would generally expect the performance of a flexible statistical learning method to be worse than an inflexible method because a flexible method would result in an overfitted model, given the small number of observations.

(c) The relationship between the predictors and response is highly non-linear. We would generally expect the performance of a flexible statistical learning method to be better than an inflexible method because more flexibility is needed for a good model fit when a greater degree of freedom and non-linearity exist.

(d) The variance of the error terms, i.e. $\sigma 2 = \text{Var}(\varrho)$, is extremely high. We would generally expect the performance of a flexible statistical learning method to be worse than an inflexible method because a flexible method would overfit to the noise characterized by the high variance of the error terms.

2. [Conceptual] Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary. Regression. Inference. n = 500; p = 3 (“profit”, “# of employees” and “industry”). Regression because CEO salary is a continuous variable. Inference gives us how the above p factors influence CEO salary.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables. (Binary) classification. Prediction. n = 20; p = 13 (“price charged for the product”, “marketing budget”, “competition price”, and ten other variables). We want to know whether a new data point will succeed or fail.

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market. Regression. Prediction. n = 52 (number of weeks in a year); p = 3 (“% change in the US market”, “% change in the British

market" and "% change in the German market"). % change in the USD/Euro is a continuous variable, not a label.

4. [Conceptual] You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response variable, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. **Real-Life Classification Application #1:**
   A bank loan officer determining whether a person is or isn't likely to default on a home loan. Response variable: Default, Yes or No. Predictors: Credit Card Debt; Income; Previous Defaults; Number of Years of Employment; Age; Credit Score; Previous Home Ownership; Previous Bankruptcy, Debt, Place of Employment; Years Employed at Current Place of Employment, Gender, Number of Children, Spouse. This is binary classification would have a goal of prediction because a lender would want to know, based on the above parameters, which potential lendee is likely to default, as determined the relationship of all or some of the predictors to the response variable.
2. **Real-Life Classification Application #2:**
   A cancer research group wants to understand from collected images of colons, whether colon tissues of mice are cancerous (dysplastic), healthy or inflamed. Response: Cancerous, Healthy or Inflamed. Predictors [include]: Fractal Box Count, # of Vessels Detected, Length Area of Vascular Crypt of Colon Wall, Diameter Measure of Vascular Crypt of Colon Wall, Fourier Characteristics of Colon Wall's Blood Vessels, Skeletonized Characteristic of Vascular Network of Colon Wall. This classification would have a goal of prediction for whether the colon of an unknown mouse is cancerous, inflamed or healthy, as determined the relationship of all or some of the predictors to the response variable. https://www.tandfonline.com/doi/abs/10.1080/00949655.2017.1374387
3. **Real-Life Classification Application #3:**
   A biogenetics research effort wants to know which bases in a primary DNA sequence occur in a nucleosome-free region. Response: Base Occurs in Nucleosome-Free Region, Yes or No. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5204302/ Predictors: # of Five-Nucleotide Sequences, # of 3-Nucleotide Combinations in One Period of a Helix, Mono-Nucleotide Distributions in DNA Fragments, Di-Nucleotide Distributions in DNA Fragments. https://academic.oup.com/bioinformatics/article/33/1/42/2525674 This classification model would have a goal of inference because it would give us understanding of the way that the response is affected by the predictors. It could potentially have a predictive goal as well, but my guess would be "not at this stage".

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. **Real-Life Regression Application #1:**
   How much revenue is a new brand extension predicted to generate over the next five years? Response: Revenue in dollar amount. Predictors: Product Cost, Customer Acquisition Cost, Price, Sales, Market Penetration, Conversion Rate and Customer Lifetime Value. This regression model would have a goal of prediction for an amount of revenue (for each item in the brand extension), given the relationship of the predictors to the response.
2. **Real-Life Regression Application #2:**
   How much of Drug A should be given to patients for the treatment of Condition A? Response: blood pressure. Predictor: dosage. This regression model would have a goal of inference for the amounts of Drug A recommended for patients on a spectrum of (grouped) blood pressure ranges.
3. **Real-Life Regression Application #3:**
   How much fertilizer and how much water should be used on a given crop in agricultural fields in order to best grow said crop? Response: crop yield. Predictor: Fertilizer Amount, Irrigated Water Amount, Rainwater Amount, Sun Amount, Heat Amount. This regression model would have a goal of inference because we want to see patterns given conditions. It could potentially be used for prediction later.

(c) Describe three real-life applications in which cluster analysis might be useful.

1. **Real-Life Clustering Application #1:**
   Which emails are spam, ham or neither, as in, which emails are compatible with my inbox and which should go to my junk folder?
   https://www.sciencedirect.com/science/article/pii/S2405844018353404
2. **Real-Life Clustering Application #2:**
   Which marketing campaign should my company use for a given demographic?
3. **Real-Life Clustering Application #3:**
   Which life insurance claims are fraudulent (are outliers)?