1. 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 predic- tors p is small.

This seems to still depend on how the data are distributed, but generally, I would say a less flexible method will perform better here, given that we have a large number of observations to average over.

(b) The number of predictors p is extremely large, and the number of observations n is small.

We might want a more flexible method in this case, since the data are sparse and we want a model that responds smoothly to possible large changes along and across predictors.

(c) The relationship between the predictors and response is highly non-linear.

A more-flexible model will clearly be expected to have better performance here, as it will reflect the non-linear nature of the real function.

(d) The variance of the error terms is extremely high.

A less-flexible function will likely respond better here, because the bias-variance trade-off is concerned with nuanced differences that are overwhelmed in a high-ε situation. The variance of $\hat{f}$ and the bias of $\hat{f}$ are insignificant compared to the variance of the error ε, so we don't gain predictability by attempting to reduce them.

2. 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.

p = 4
n = 500
This is a regression problem, as we're predicting numerical values using numerical values. Prediction is interesting here, because we want to be able to predict CEO salary as a function of the predictors we find significant of the 4 available.

(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 prod- uct we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

p=14
n=20
Another prediction problem, because we're interested in a predicted outcome -- success or failure -- as a function of the various predictors. This could be considered semi-categorical, since at least one predictor has a classification nature, but I would say it is a classification problem because the goal is to predict a class: failure or success.

(c) We are interesting in predicting the % change in the US dollar 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 dollar, the % change in the US market, the % change in the British market,

and the % change in the German market.

n=52
p=4
A clear regression setting, but this is an inference problem, not a prediction problem. With inference, we have a starting place and attempt to predict the change in a variable as a function of other observed rates: in this case, we have a known US dollar price, and we want to predict how it will change given rate shifts in other markets, so inference clearly applies.

4. 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, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Image identification. The predictors could be things like "distribution of greyscale intensity", "distribution of colors", and any number of clever things I'm sure machine learning professionals have thought up. The response is the most probably classification. This is a prediction.

Galactic classification. Really this is very similar to general image identification, but we classify galaxies using very specific spectral bands for the predictors that involve light intensity, but then we also look at how strong particular spikes or dips in the spectrum are, so we might have predictors for "emission line strength" for several spectral features. The response is the most likely galactic classification. This is a prediction.

Speech recognition. The predictors would perhaps be the audio spectrum, with the response being the word the audio spectrum corresponds to. This would predict the most likely word for the audio received.

(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.

Marketing data is obvious. The predictor is perhaps how much was spent on a certain type of marketing, or a few types of marketing -- this is now sounding like the example from the book. The response is an amount sold for the same fiscal period. You could use inference or prediction here: inference to how how many addition sales you might add by spending marketing funds, or prediction by asking just "how many sales did we see when we spent X amount on marketing?"

I want to try to use this for my project: understanding the time delay, or reverberation, of a dynamic spectral feature compared against a similarly dynamic reference feature. 2 predictors, line-of-sight velocity and time delay, give a response of light intensity. Our task is to predict the light intensity as a function of these predictors. This is actually a vanguard question in astrophysics, and I'll bet somebody is already trying to do this!

Maybe something municipal. I could predict the taxable income of a city based on a number of predictors, like availability of mass transit or highways, demographics, resources, distance to neighbouring cities, and all kinds of things, then the response would continue to just be taxable income given all of these inputs. Perhaps it would be good to consider an inference questions here, for example: how would my city's taxable change if I increased the availability of public transit?

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

Categorizing star type by spectral band strengths.

Plant and animal species identification.

Tracking objects in sensor data.

9. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are quali- tative?

mpg, horsepower, weight, acceleration, and displacement are all clearly quantitative.

cylinders I think is arguably qualitative because each number of cylinders defines a somewhat broad class of vehicles. For the years, the same argument might apply: each year is a class of vehicles. The origin is clearly qualitative, and so is name.

(b) What is the range of each quantitative predictor? You can an- swer this using the range() function.

```
$mpg
[1]  9.0 46.6

$cylinders
[1] 3 8

$displacement
[1]  68 455

$horsepower
[1]  46 230

$weight
[1] 1613 5140

$acceleration
[1]  8.0 24.8

$year
[1] 70 82
```

(c) What is the mean and standard deviation of each quantitative predictor?

```
$mpg
      mu    sigma
23.445918  7.805007

$cylinders
      mu    sigma
5.471939 1.705783

$displacement
    mu   sigma
194.412 104.644

$horsepower
    mu     sigma
104.46939  38.49116

$weight
    mu     sigma
2977.5842  849.4026
```

```
$acceleration
    mu     sigma
15.541327  2.758864

$year
    mu     sigma
75.979592  3.683737
```

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
$mpg
       mu     sigma
24.404430  7.867283

$cylinders
      mu    sigma
5.373418 1.654179

$displacement
    mu     sigma
187.24051  99.67837

$horsepower
    mu     sigma
100.72152  35.70885

$weight
    mu     sigma
2935.9715  811.3002

$acceleration
    mu     sigma
15.726899  2.693721

$year
    mu     sigma
77.145570  3.106217
```

I've now changed my mind and say that both year and cylinders are quantitative, since there is plenty of sense about talking about the mean and std in those predictors for this set of data.

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

I'll just make all the graphs, included at auto_pairs.png. There are a number of uncorrelated predictors, it seems, but many relationships can also be discerned. Mpg and cylinders; mpg and displacement; mpg and horsepower; mpg and weight; mpg and year, even; horsepower and displacement; really, there are many relationships, but the interesting ones are probably with the mpg. The strong linear relationships between horsepower, weight, and displacement make sense because they're pretty much correlated by design, as engineers make larger engines to handle more weight and so on. The relationships between this overall trend, is that as they increase, mpg decreases. We also see that mpg increases as the year increases, i.e., as we develop more sophisticated technology.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg ? Justify your answer.

Well, I pretty much just answered that. The year is a great predictor: it appears we will likely continue to improve mpg slowly and in a

linear fashion with time. There is a non-linear relationship that gives a strong mpg response as weight/displacement/horsepower decrease, so it's quite clear that these are a strong predictor of mpg. There's also a relationship with cylinders, but again, this is really just part of the trend of vehicles with more weight being designed with larger engines. Finally, it also seems that origin "3" makes cars with slightly better gas mileage than origin "2" and again 2 makes cars with better mpg than origin "1". I can't find it in the text, but I assume origin 3 is Japan, 2 is Europe, and 1 is US, just based on my own personal bias about society.