Econ 573 **Juan Alvarez**
Prof. Andrii Babii **January 20th, 2025**

Part 1

Ex. 1, 2, 4, 5, 6 from Chapter 2 of ISL

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 predictors p is small.

A flexible statistical learning method would perform better than an inflexible method. The extremely large sample size reduces the higher variance associated with more flexible models, mitigating the risk of overfitting and allowing them to capture possible complexities in the true relationships.

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

A flexible statistical learning method would perform worse than an inflexible method. Due to the limited data, an inflexible model would keep the prediction to include unnecessary noise arising from the presence of unnecessary predictors. A flexible model would be more likely to overfit the data.

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

A flexible statistical learning method would perform better than an inflexible method. Flexible models don't assume a specific functional form for the relationship between predictors and the responses, allowing them to capture highly non-linear relationships between them.

(d) The variance of the error terms, i.e. $\sigma 2 = Var(\epsilon)$, is extremely high.

A flexible statistical learning method would perform worse than an inflexible method. The variance of the error terms is extremely high alluded to the fact that the data contains a lot of noise. A flexible model would be prone to wrongfully overfit the data due to the presence of this noise. An inflexible would impose a structure on the data to reduce the variance.

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.

Econ 573 **Juan Alvarez**
Prof. Andrii Babii **January 20th, 2025**

(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.

This scenario is a regression problem with inference as its primary interest. We are interested in explaining CEO salary, a continuous variable, in terms of the number of employees, industry for each firm, that is how each of these explanatory variables influence CEO salary. The sample size, or n, is 500 due to the fact that we are collecting data on the top 500 firms in the US. The number of parameters, or p, is 3, because there are 3 factors of interest when attempting to explain CEO salary for each firm: profit, number of employees, and industry.

(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.

In this scenario, we are looking to forecast how the new product will be received and classify the results as one two categories: success or failure. So, it is a classification problem with prediction as its primary interest. The sample size, or n, is 20 as we are collecting data on 20 similar products that were previously launched. The number of parameters, or p, is 13 because we are looking to predict whether the new product will be a success or failure based on price charged for the product, marketing budget, competition price, and ten other variables.

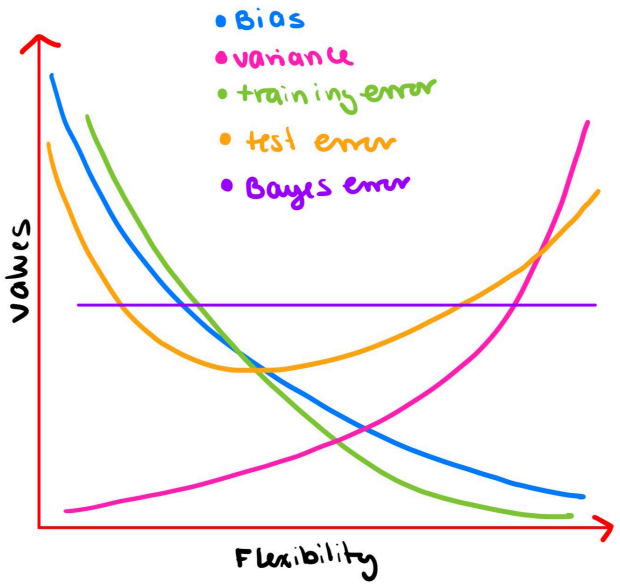
(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

In this scenario, we aim to predict percent changes in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. So we can categorize it as a regression problem with prediction as its primary interest. Because we are collecting weekly data for all of 2012 and there are 52 weeks in a year, the sample size, or n, is 52. The number of parameters, or p, is 3 due to the fact that the percent changes in the USD/Euro exchange rate will be explained by the percent change in the US market, the percent change in the British market, and the percent change in the German market.

3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical

learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(b) Explain why each of the five curves has the shape displayed in part (a).

The typical (squared) bias decreases as the flexibility of a method increases. More inflexible methods force simplified models onto the data, which might not capture the true relationships, while flexible methods reflect the true relationships in the data in a more accurate way. Inflexible methods have low variance because moving any single observation will likely only cause a small shift in the position of the curve, while the opposite is true for flexible methods. So, the variance of a method increases with its flexibility. When using some training data, flexible methods will always always fit the simulated data, causing the errors to decrease with flexibility and giving rise to overfitting issues. However, when using test data, we observe that errors are high for inflexible models but decreasing as flexibility increases until it reaches a minimum value, and then they start increasing with flexibility. The Bayes Error is a constant term because it does not depend on x or the flexibility of the method.

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.

One real-life application is constructing a model by classifying high and low mobility rates in a specific city based on income, age, and location. high/low mobility rates would be the response, while income, age, and location are the predictors. The goal of this application is inference, as we aim to explain how these factors influence high versus low mobility rates. Another example is a model that predicts after-graduation success for students across different time frames such as long run and short run based on predictors such as GPA, industry of work, and academic program. A third example is whether international alumni from a specific university choose to stay in the US and work, or work outside of the USA. This model could attempt to explain this decision based on country of origin, academic program, and parent's education.

(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.

An application in which regression is useful is a model to explain wages (response) based on predictors such as age, years of experience, education level, and location. The goal for this model is prediction. Another application is to build a model to predict the price of a new product based on predictors such as the competitors' prices, the cost of production, and marketing budget. A third example is to

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

Cluster analysis can be useful when grouping customers into different segments based on their buying behavior so firms are able to maximize profit by using targeted marketing. Governments can use clustering analysis to differentiate between geographic areas with different socio-economic needs and provide needed support. Tiktok could use cluster analysis to group users based on their activity in the app, such as influencers, casual users, etc.

5. *What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?*

Less flexible approaches for regression or classifications are more likely to be biased, but are low in variance. However, their results are easier to interpret, allowing for simple relationships between response variables and their predictors. Conversely, more flexible approaches have less bias but higher variance. Flexible methods also face the problem of overfitting the data, but are more adequate for explaining complex relationships. When inference is the main goal, inflexible models make for the better models due to their easy interpretability. Understanding how individual predictors are associated with the response becomes simpler. In the cases where

prediction is the main interest, it might be best to use flexible models. Flexible methods are more likely to give accurate predictions .

6. *Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?*

A parametric approach entails making an assumption about the functional for, or shape, of f. Then we create a procedure that uses training data to fit the model by creating parameters that are estimated. However, this model usually does not match the true form of the function, which could lead to poor estimations. Conversely, a non-parametric approach does not make explicit assumptions about the functional form of, and instead it seeks an estimate of the function that gets as close to the data points as possible. By avoiding the parametric approach's initial assumptions, there is potential to accurately fit a wider range of possible shapes for f. However, non-parametric approaches require a very large number of observations in order to obtain an accurate estimate for f.

Part 2

Ex. 8, 9, 10 Chapter 2 of ISL

Tip: You can generate beautiful PDF/HTML files with your data analysis from R notebooks if you click the Knit button at the top of the notebook. You should submit the output as a part of your solution

Part 3

See benjerry_start.R for code to get you started.

1. Explore the data and visualize: what variables are interesting? Choose a few, plot them together, and tell a story.
2. Describe the regression model in the code. Improve it?
3. Take the p-values from your regression and look for evidence of association. Relate what you learn to your story from 1