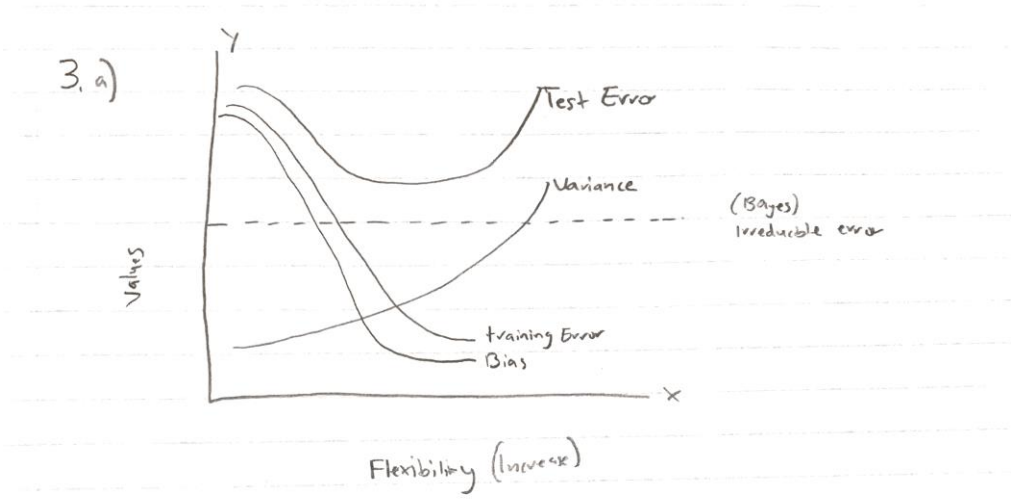
Siddhant Rai

Chapter 2

3.

a)



b) Bias and variance are inversely related in this meaning as flexibility increases , these is an increase in variance, and bias decreases. This is because Bias is used to simplify a complex regression while Variance estimates how close our training data set is to the model how our estimated f^(x) of we change a single point on the data set. For training vs test error, the training data decreases, and test initially decrease before increases again as flexibility increases. When we are given a small training & large test dataset, we to the overfit the model because the patterns in the train may not be in the test. And with the irreducible error(Bayes), it is horizontal and constant, and the test error never crosses it

c) The advantages of a flexible approach for regression or classification are that we can reduce bias and estimate a larger number of parameters in a nonlinear model. However, the disadvantages are if the parameters are too close to the error this can cause overfitting, in the case of flexible approach, it is preferrable if we want to the interpret the model's parameters, but in the case of less flexible, we only want the prediction.

10.

a) There are **506 rows & 14** columns in the dataset. The rows represent the value, and the columns represent the labels for each value.

b) The findings based on my pair plot among all columns. There is a **positive correlation** between RM & MDEV. **Negative correlation** between LSTAT & RM, NOX & DIS, and LSTAT & MDEV.

c) The is a between crime rate per capital and the following predictors NOX (0.42), RAD(0.62),
   i. **Relatively and significantly positive correlation** NOX (0.42), RAD (0.62), TAX (0.58), and LSTAT (0.45).
   ii. **Slightly positive correlation** AGE (0.35), PTRATIO (0.29), INDUS (0.4)
   iii. **Marginally negative** CHAS (-0.055)
   iv. **Relatively Negative correlation** Zin (-0.2), RM (-0.22), DIS (-0.38), B (-0.38),MDEV (-0.39)

d) For this problem I did three boxplots
   i. The Crime rate per capita boxplot is skewed mostly right and shows there is a **low rate** of crime between 0-10 range on the boxplot. There is however a large rate of outliers close to the Largest non-outlier, while there are approximately 8 suburbs who have a crime rate in the outlier range above 40.
   ii. Tax(full-value property-tax rate per $10,000) boxplot is normally distributed, slightly left skewed, but overall, no outliers, median of 350 and range between 200-700 so no extreme high or low tax rates.
   iii. Ptratio (pupil-teacher ratio by town) boxplot is skewed toward the left and has a median of 19, median 19, range 12-22, there are two outliers below the range, but not a large amount, and overall, no extreme high or low ratios.

e) **35** suburbs bound the Charles River.

f) The median pupil-teacher ratio among the towns is **21.2**.

g) The suburb with the lowest median value of the owner-occupied homes has a **MDEV value of 5.**

The other predictors compared to MDEV are:		Range of the predictors:
CRIM	**0.0136**	88.96988
ZN	**0.0000**	100.00000
INDUS	**1.6900**	27.28000
CHAS	**0.0000**	1.00000
NOX	**0.3850**	0.48600
RM	**4.1380**	5.21900
AGE	**18.5000**	97.10000
DIS	**1.1370**	10.99690
RAD	**1.0000**	23.00000
TAX	**188.0000**	524.00000
PTRATIO	**14.7000**	9.40000
B	**0.3200**	396.58000
LSTAT	**5.5700**	36.24000

Compare to the Range , all these values **are way below the range** in each predictor except for R M & NOX which are the closest to the range. There are also values of 0 for ZN and CHAS, 0.32 f or B, 1 for RAD, the min for the predictors in both columns.

h) There are **64 suburbs** which have 7 or more rooms per dwelling & **13 suburbs** which have 8 or more rooms per dwelling. For the 13 suburbs, the people who live in these homes have a mean AGE of 71.5, and low crime rate meaning these are more than likely nursing homes for older individuals.

Chapter 3

3.

a) $y = 50 + 20\ GPA + 0.07\ IQ + 35\ Gender + 0.01\ GPA*IQ - 10\ GPA*Gender$

Gender: {Male:0, Female:1}

Males: 50 +20 GPA +0.07 IQ +0.01 GPA*IQ

Females: 50 +20 GPA +0.07 IQ +35 + 0.01 GPA*IQ – 10 GPA

Subtract: Females = 35 -10 GPA -> GPA =3.5

iii) is correct because in the equation above when we subtract at a GPA average of 3.5 or more, women earn less than men.

a) Female: 50 +20 GPA +0.07 IQ +35 + 0.01 GPA*IQ – 10 GPA
50 +20(4) +0.07(110) +35 + 0.01 (4*110) – 10(4) **= 137.1 -> $137,100**

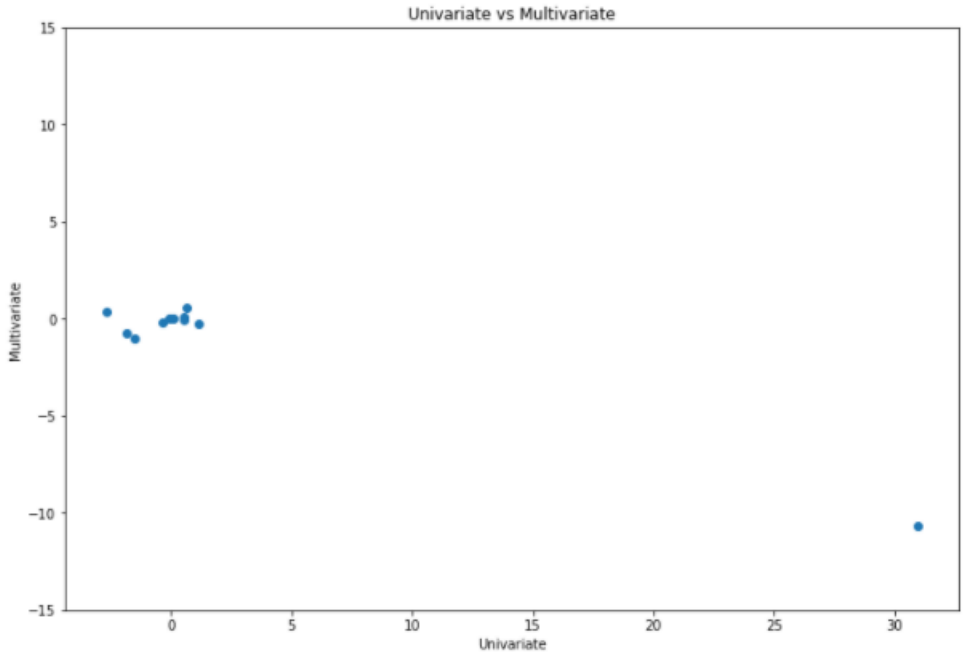
b) **False**, because even if the interaction is small, if the p value is below 0.05 the coefficient is statistically significant.

15.

a) **Every predictor except Charles River is statistically significant**. When looking at the graphs, the CHAS dummy variables is completely separated and has no normal distribution or correlation. But overall, the predictors are far away from fitted regression line when looking at other predictor's correlation with crime.

b) We can reject the null hypothesis for **RAD, DIS,ZN, NOX and MDEV** because they have a p-value below 0.005.

c)



d) There is nonlinearity for the following variables **INDUS, NOX, AGE,DIS, PTRATIO, MDEV**