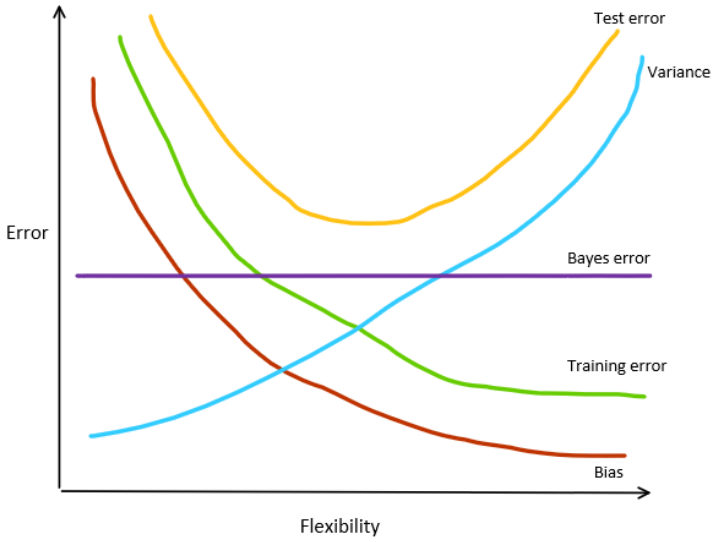
Jayoung Kang

Problem Set 1

1. *Chapter 2*
   i. *Question 3:*
      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).*
         - **Bias**: Bias measures the difference in model prediction and the true value. As flexibility increases, the bias decreases because it will be fitting the model more to the given data.
         - **Variance**: Variance measures how much the prediction would change if the training set were changed. Variance is high when flexibility is low because the fitted model is independent of the data. As flexibility increases, variance will increase because the model becomes more dependent on the data.
         - **Training error**: Training error is high when flexibility is low and decreases with flexibility because it measures the average difference between the prediction and the observations.
         - **Test error**: Test error is composed of bias, variance and Bayes error. Since bias and variance move in opposite direction when flexibility increases and Bayes error is constant, the addition of these components gives the convex parabola shape and the minimum is the sweet spot of the bias-variance tradeoff.
         - **Bayes error**: Bayes error is constant and does not change with flexibility because it is the error that is not dependent on the input (x) variable.

ii. *Question 5*

a. *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?*

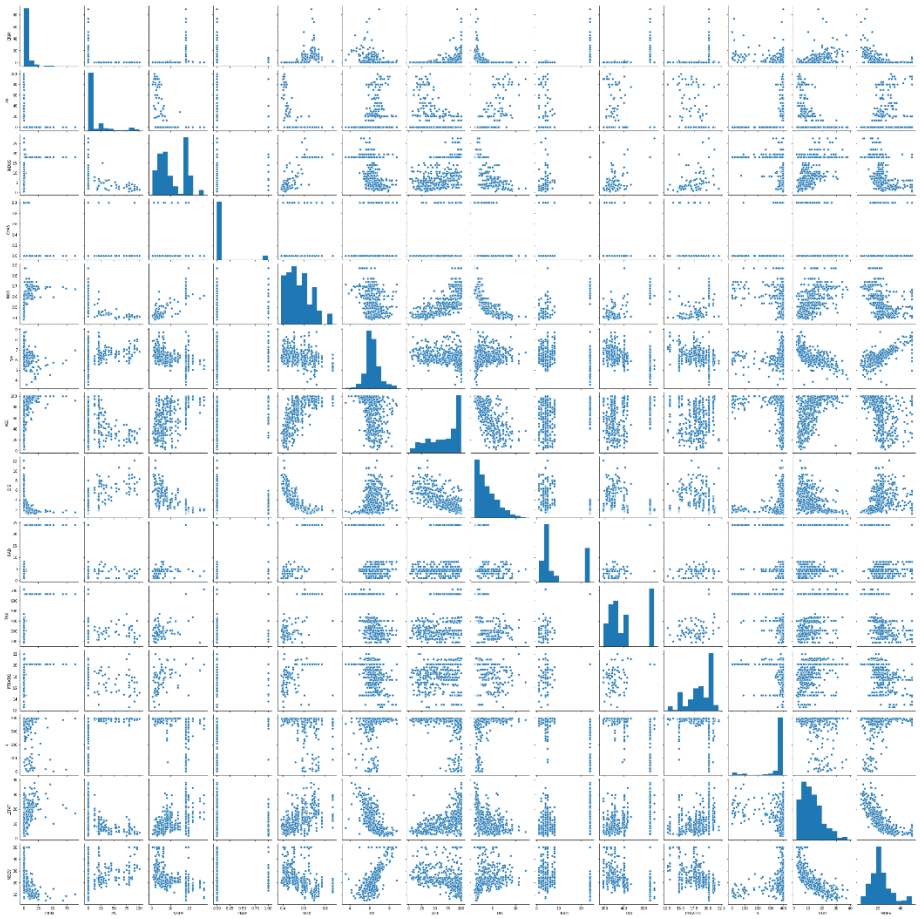
- The advantages and disadvantages of flexibility of the model is related to the tradeoff between flexibility and interpretability. The disadvantage of more flexible models is that they tend to overfit and so it is more difficult to generalize the relationship between X and Y. Less flexible models are more interpretable and therefore it would be more preferred to use a less flexible approach if the main goal is to gain more interpretability from the model. An advantage of more flexibility is that it has higher predictive ability. Therefore, it is more preferred to use flexible models when we are building algorithms where we care more about accurate prediction than the interpretability.

iii. *Question 10*

a. *How many rows are in this data set? How many columns? What do the rows and columns represent?*

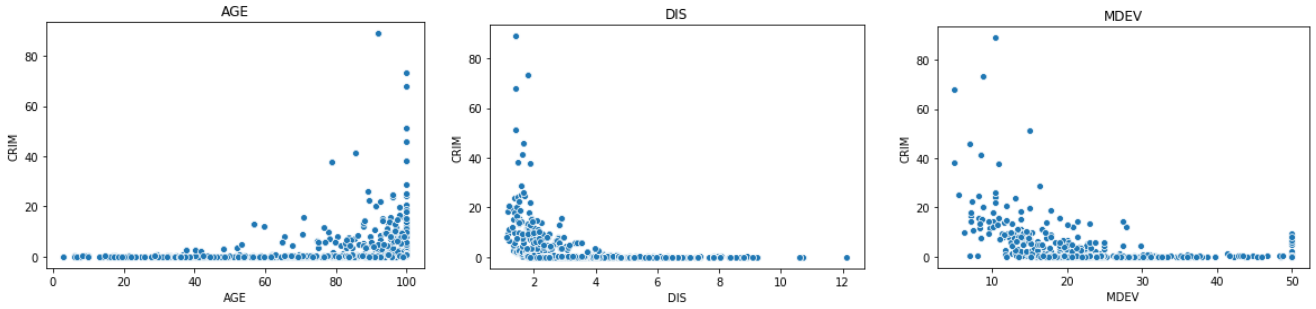
- 506 rows and 14 columns. The rows represent the 506 cases of housing data in Boston and the columns are the attributes of each case such as per capita crime rate and nitric oxide concentration.

b. *Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings*

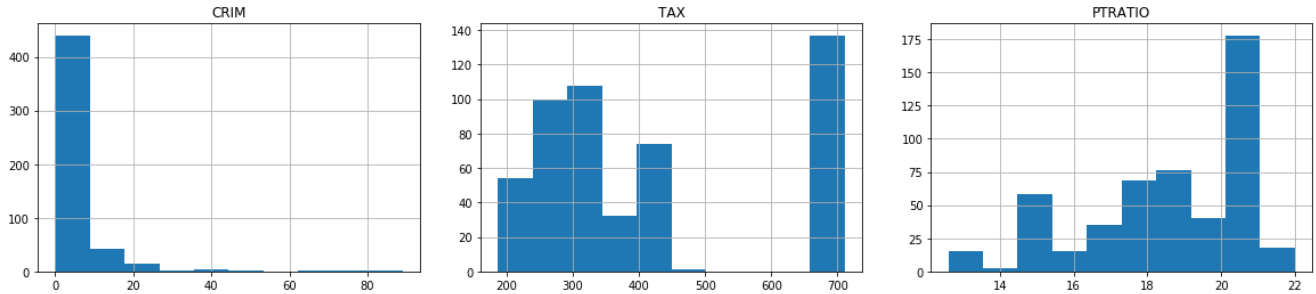


- We can observe some correlation between the variables. Below are some correlations that seem apparent from just looking at the pairwise plots. However, these would need to be corroborated by actually running a model.
- CRIM: positively correlated with AGE, negatively correlated with DIS, MDEV
- ZN: positively correlated with DIS, negatively correlated with INDUS, NOX, LSTAT
- INDUS: positively correlated with NOX negatively correlated with DIS
- AGE: positively correlated with NOX
- DIS: negatively correlated with NOX, AGE
- LSTAT: negatively correlated with RM, MDEV

c. *Are any of the predictors associated with per capita crime rate? If so, explain the relationship.*
- Crime rate is positively correlated with AGE and negatively correlated with DIS, MDEV. However, the relationship does not look linear and suggests the possibility of a non-linear exponential or quadratic relationship.
- It seems to make sense that areas with higher crime rate is associated with older houses and it also makes sense that areas with higher crime rate is associated with lower housing values and smaller distance to metropolitan areas with employment centers



d. *Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.*
- **CRIM**: most suburbs (400+) have low crime rate around 0 and 10 but there is a long tail of suburbs with particularly high crime rate of between 20 and 80
- **TAX**: there is a clear divide between those with tax rate between a group around 200 and 500 and another group with particularly high tax rate of around 650 and 700
- **PTRATIO**: the distribution is generally skewed to the left and there is a high number of suburbs (175+) that have ptratio of above 20 but they do not seem particularly high relative to the general distribution of between 12.6 to 22



e. *How many of the suburbs in this data set bound the Charles river?*
   - 35 suburbs

f. *What is the median pupil-teacher ratio among the towns in this data set?*
   - 19.05

g. *Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.*
   - Suburbs 398 and 405 have the lowest MDEV

	398	405	Comparison to overall ranges of predictors
CRIM	38.3518	67.9208	both above 75th percentile
ZN	0	0	both at min
INDUS	18.1	18.1	both at 75th percentile
CHAS	0	0	both at min
NOX	0.693	0.693	both above 75th percentile
RM	5.453	5.683	both below 25th percentile
AGE	100	100	both at max
DIS	1.4896	1.4254	both below 25th percentile
RAD	24	24	both at max
TAX	666	666	both at 75th percentile
PTRATIO	20.2	20.2	both at 75th percentile
B	396.9	384.97	at max for 398 and between 25th and 50th percentile for 405
LSTAT	30.59	22.98	both above 75th percentile
MDEV	5	5	both at min

h. *In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.*
   - More than 7 rooms: 64
   - More than 8 rooms: 13
   - Suburbs with more than average 8 rooms per dwelling have lower average CRIM (crime rate), higher average of MDEV (median value of owner occupied homes) and lower average of LDEV (percentage of lower status population than the total data average)

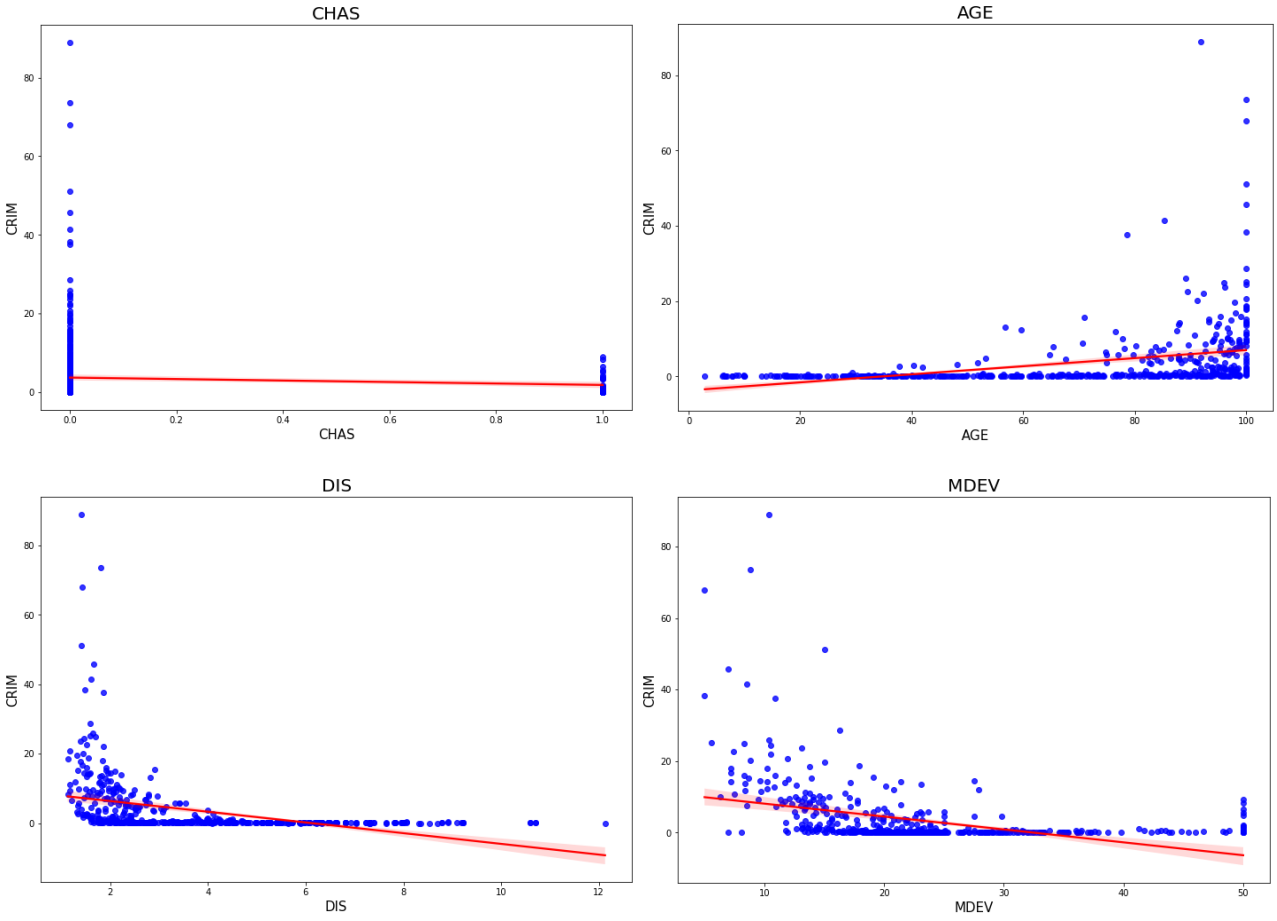
2. *Chapter 3*
   i. *Question 3*
      a. *Which answer is correct, and why?*
         - (iii) The FOC when deriving the model by gender shows 35-10*GPA meaning that if GPA is high enough, women earn less than men on average.

      b. *Predict the salary of a female with IQ of 110 and a GPA of 4.0*
         - 50 + 4*20 + 110*0.07 + 1*35 + 0.01*4*110 - 10*4*1 = 137.1
         - She is expected to earn 137.1k dollars

c. *True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.*
   - False. Evidence for interaction effect is measured by the t-statistics or the p-value.

ii. *Question 15*

a. *For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.*
   - All predictors except the CHAS have statistically significant association between the predictor and the response because they have a p-value of below 0.005.
   - The plot of CHAS shows almost a straight regression line suggestive that CRIM and CHAS do not have statistically significant association.
   - The three variables mentioned in question 10c) (AGE, DIS, MDEV) all show some statistically significant association.



Predictor	Correlation
ZN	Yes, negative
INDUS	Yes, positive
CHAS	No, p-value=0.214

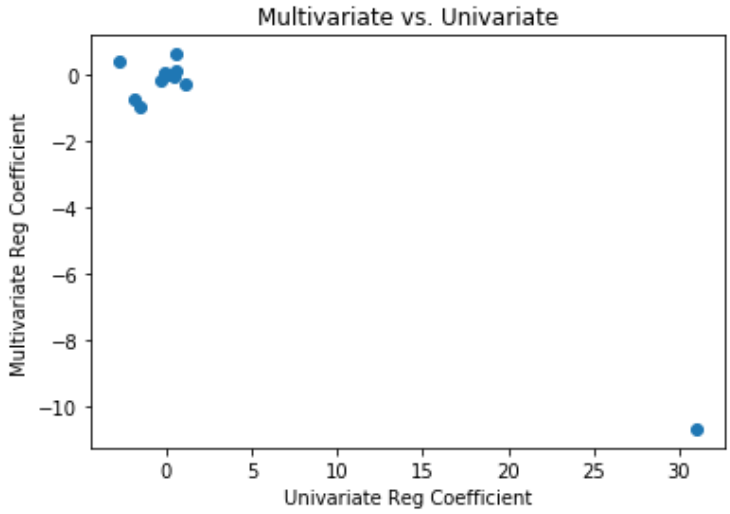
NOX	Yes, positive
RM	Yes, negative
AGE	Yes, positive
DIS	Yes, negative
RAD	Yes, positive
TAX	Yes, positive
PTRATIO	Yes, positive
B	Yes, negative
LSTAT	Yes, positive
MDEV	Yes, negative

b. *Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis*

- The F-statistic is 30.73 meaning that the variables are jointly significant and R-squared is 0.448 meaning that the predictors explain 44.8% of the variability in the crime rate.
- The predictors that reject the null are **ZN, NOX, DIS, RAD** and **MDEV** because the p-values are below 0.005 and the T-stats each have absolute values larger than 2.

c. *How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.*

- There are fewer significant variables in (b). But the variables significant in (b) are also significant in (a). It is possible there are fewer significant variables in the multivariate case due to multicollinearity.
- The plot below shows that NOX is the only variable that shows a large difference between its multivariate and univariate coefficient.



d. Is there evidence of non-linear association between any of the predictors and the response?

- There is evidence of non-linear association for all predictors except CHAS, RM, RAD, B, LSTAT and TAX where neither the squared or cubed variable were significant. All other predictors showed a statistically significant association for either the squared or cubed variable or both.