STATS 202: Data Mining and Analysis Homework #1

Adam Kainikara

July 7, 2023

Problem 1 (4 Points)

Chapter 2, Exercise 2 (p. 52).

a) This problem is a regression problem as our output variable is quantitative and continuous. The output variable is the CEO's salary we are investigating which factors affect the CEO's salary. In this problem we are more interested in inference. Inference problems are those where we are interested in understanding how differences in variables might affect Y. Prediction problems involve using the variables to help predict Y, when we can not obtain Y. In this problem we have variables such as CEO salary, profit and number of employees, and we aim at determining how these variables affect the output variable, CEO salary, rather than trying to predict what the CEO's salary is. In this problem n = 500 and p = 3 because n is the number of observations and p is the number of variables. We have 500 observations because we asked 500 firms and 3 variables because we record profit, number of employees, and industry.

b) This problem is a classification problem as our output variables, success or failure is qualitative and discrete. This problem is a prediction problem as we are using many variables from similar products such as price charged for a product and marketing budget to predict whether a new product would succeed or fail. In this problem n = 20 and p = 13 because we looked at 20 different observations (products) and looked at 13 variables including price of the product and marketing budget.

c) This problem is a regression problem as our output variable, % change in USD/Euro exchange rate, is quantitative and constitution. The problem is a prediction problem as we are using multiple different market's data (variables) to help us predict the % change in USD/Euro exchange rate. In this problem n = 52 and p = 3 because the problem states that we took weekly data and collected data from 3 different markets.

Problem 2 (4 Points)

Chapter 2, Exercise 3 (p. 53).

a) See inserted graph Figure

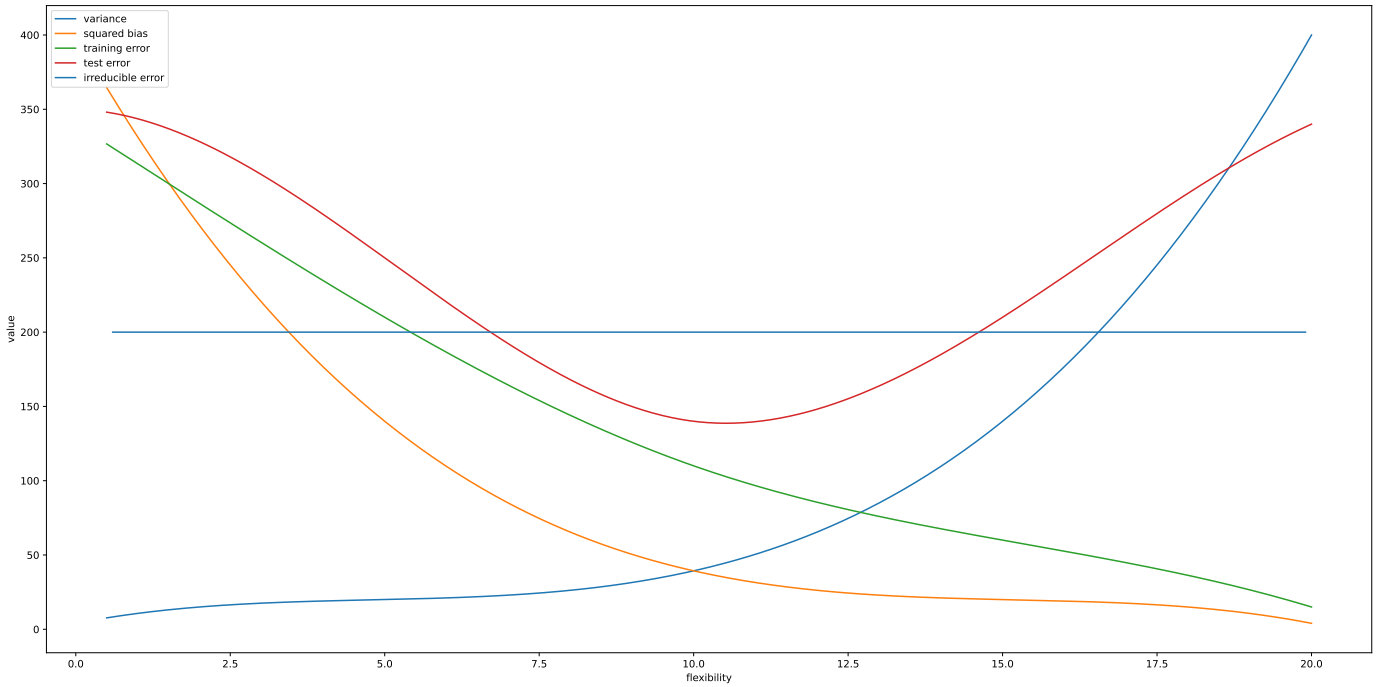


Figure 1: Variance, Squared Bias, Training Error, Test Error, Irreducible Error: Problem 2

b) Training error decreases as flexibility increases because flexibility enables the function to be fitted more accurately. With more data fitted, the mean square error decreases as flexibility increases. Irreducible error is a flat line because that will never go away and is a constant amount. As flexibility increases variance increases and bias decreases. This relationship can be seen through bias variance trade off.

Problem 3 (4 points)

Chapter 2, Exercise 7 (p. 54).

a) The Euclidean distance can be found by square rooting the sum of the differences between each component of the observations and test point.

Obs.	$X1$	$X2$	$X3$	Y	$Euclidean\,Distance$
1	0	3	0	*Red*	$\sqrt{(0-0)^2+(3-0)^2+(0-0)^2}=3$
2	2	0	0	*Red*	$\sqrt{(2-0)^2+(0-0)^2+(0-0)^2}=2$
3	0	1	3	*Red*	$\sqrt{(0-0)^2+(1-0)^2+(3-0)^2}=3.2$
4	0	1	2	*Green*	$\sqrt{(0-0)^2+(1-0)^2+(2-0)^2}=2.2$
5	-1	0	1	*Green*	$\sqrt{(-1-0)^2+(0-0)^2+(1-0)^2}=1.4$
6	1	1	1	*Red*	$\sqrt{(1-0)^2+(1-0)^2+(1-0)^2}=1.7$

b) This is computing k-nearest neighbors to the test point (0,0,0) with k = 1. The nearest observation to the test point is observation 5 which has a Euclidean distance of about 1.4. Because observation 5 is green, we predict that the test point will be green.

c) This is computing k-nearest neighbors to the test point (0,0,0) with k = 3. The three nearest observations to the test point are observations 5, 6, and 2. These 3 observations have a Euclidean distance of about 1.4, 1.7, and 2.2 respectively. Observation 5 is green while observations 2 and 6 are red. Because red is the majority, we predict that the test point will be red.

d) The best value for K in this problem would be small. This is because of the decision boundary is non linear. For example if the decision boundary was very curvy the graph would show a lot of variance. Due to the high variance, a small value of K would be best for this problem.

Problem 4 (4 points)

Chapter 12, Exercise 1 (p. 548).

a)

$\frac{1}{|C_k|}\Sigma_{i,h\in C_k}\Sigma_{j=1}^{p}(x_{ij}-x_{hj})^2 = 2\Sigma_{i\in C_k}\Sigma_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$

Addition is commutative so we can re-arrange the sigmas

$\frac{1}{|C_k|}\Sigma_{j=1}^{p}\Sigma_{i,h\in C_k}(x_{ij}-x_{hj})^2 = 2\Sigma_{j=1}^{p}\Sigma_{i\in C_k}(x_{ij}-\bar{x}_{kj})^2$

We will expand $\Sigma_{i\in C_k}(x_{ij}-\bar{x}_{kj})^2$, and just because it is easier to type temporarily $\bar{x}_{kj}=\mu$, $|C_k|=n$ and $i^{'}$ from the problem is really hard to see it has been renamed to h.

$$\begin{array}{ccccc}
(x_1-\mu)^2 & +(x_2-\mu)^2 & +\cdots & +(x_n-\mu)^2 \\
x_1^2-2x_1\mu+\mu^2 & +x_2^2-2x_2\mu+\mu^2 & +\cdots & +x_n^2-2x_n\mu+\mu^2 \\
= & \Sigma x_i^2 & -2\mu\Sigma x_i & +n\mu^2 \\
 & =\Sigma x_i^2 & -2n\mu^2 & +n\mu^2
\end{array}$$

Because $\Sigma x_i = n\times\mu$

$2\times \quad (\Sigma x_i^2 \quad -n\mu^2) \quad = 2(\Sigma x_i^2 - n\mu^2)$ (Added scale factor of 2 back from the equation)

Next we expand $\Sigma_{i,h\in C_k}(x_{ij}-x_{hj})^2$

$$\begin{array}{ccccc}
(x_{1j}-x_{1j})^2+ & (x_{1j}-x_{2j})^2+ & \cdots & +(x_{1j}-x_{nj})^2+ \\
(x_{2j}-x_{1j})^2+ & (x_{2j}-x_{2j})^2 & +\cdots & +(x_{2j}-x_{nj})^2+ \\
 & \vdots & & \\
(x_{nj}-x_{1j})^2+ & (x_{nj}-x_{2j})^2+ & \cdots & +(x_{nj}-x_{nj})^2 \\
=x_{1j}^2-2x_{1j}x_{2j}+x_{2j}^2 & +x_{1j}^2-2x_{1j}x_{3j}+x_{3j}^2 & +\cdots & +x_{1j}^2-2x_{1j}x_{nj}+x_{nj}^2 \\
= & nx_{1j}^2+ & \Sigma x_{ij}^2 & -2x_{1j}\Sigma x_{ij} \\
nx_{2j}^2+ & \Sigma x_{ij}^2 & -2x_{2j}n\mu & \\
 & \vdots & & \\
=n\Sigma x_{ij}^2 & +n\Sigma x_{ij}^2 & -2n\mu\Sigma x_{ij} & \\
=2n\Sigma x_{ij}^2 & -2n^2\mu^2 & & \\
=2n(\Sigma x_{ij}^2 & -n\mu^2) & & \\
\frac{1}{n}\times & 2n(\Sigma x_{ij}^2-n\mu^2) & =2(\Sigma x_{ij}^2-n\mu^2) &
\end{array}$$

We have shown that the left and right sides are the same for a particular j adding the Σ_j back into the original problem is just adding p copies of equal terms to the left and right hand sides. Therefore the identity is proven.

b) Objective 12.17 aims at minimizing the average squared distance between each observation. K means clustering algorithm uses a repetitive process of computing and moving the cluster centroid and assigning each observation to a cluster based on distance until the centroids of new clusters are not changing. The process of changing the centroid reduces g the distance between each observation in the cluster and the centroid. K means clustering algorithm decreases Objective 12.17 in each iteration. In the first step of each iteration,

the relocation of the centroid reduces the sum of the distances for each point, and the mean. Doing this process over and over can only reduce the distances.

Each iteration, the centroid of each cluster moves to the mean position of all observations in the cluster. The next step involves relocating observations to the closest centroid which reduces the mean squared value.

Problem 5 (4 points)

Chapter 12, Exercise 2 (p. 548).

See attached Figure2

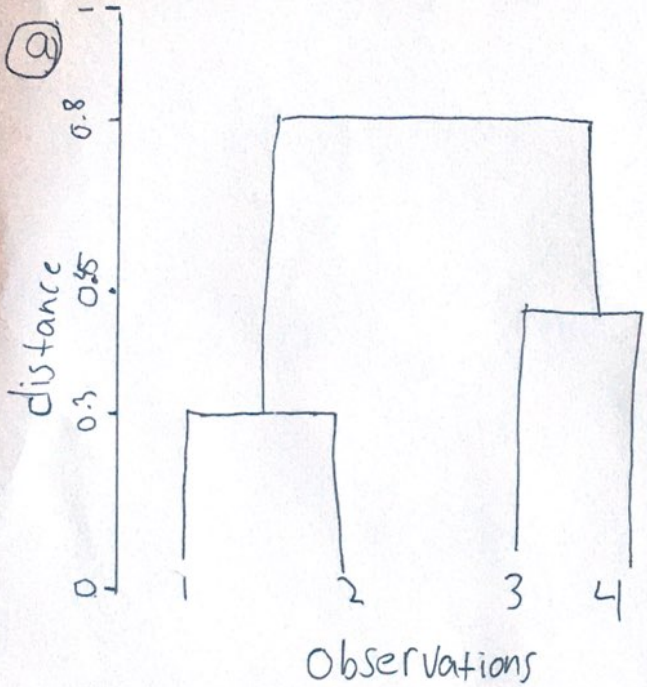
Stats 202 HW#1

Adam Kainikara

Problem 5 (4 points)

Chapter 12 Exercise 2 (p.548)

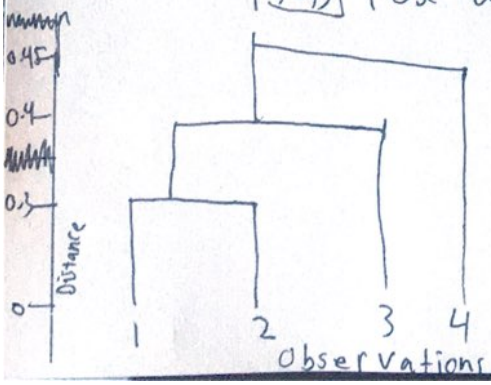
(2) (a)



(b) 1,2 fuse at 0.3

[1,2] fuse with 3 @ 0.4

[1,2,3] fuse with 4 at 0.45



This is 'D'

(d) If the dendogram is cut so two clusters result, one cluster would have 1, 2, 3 and the other cluster would have 4

This is 'C'

(c) If the dendogram in 'a' was cut so that two clusters result, one cluster would have 1, 2 and the other cluster would have 3, 4

(E)

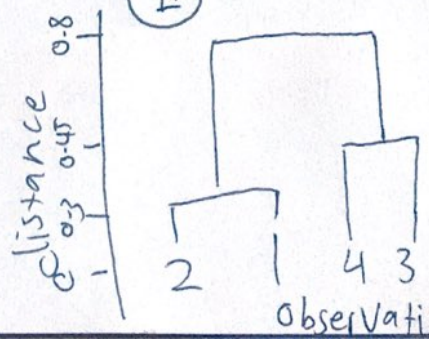


Figure 2: Problem 5

Problem 6 (4 points)

Chapter 12, Exercise 4 (p. 549).

a) Complete linkage uses the farthest two data points and single linkage uses the two closest data points. Due to complete linkage using the farthest distance, the fusion would generally occur higher on the tree while single linkage would occur lower on the tree. If the distances between the two clusters were the same, it is possible that the fusions would occur at the same point. Because the distance information is not provided, there is not enough information to tell which fusion would occur higher on the tree.

b) Because there are only two points, 5 and 6, the type of linkage (single or complete) is not relevant to this problem. The height will be the same.

Problem 7 (4 points)

Chapter 12, Exercise 9 (p. 550).

a) See inserted Figure 3

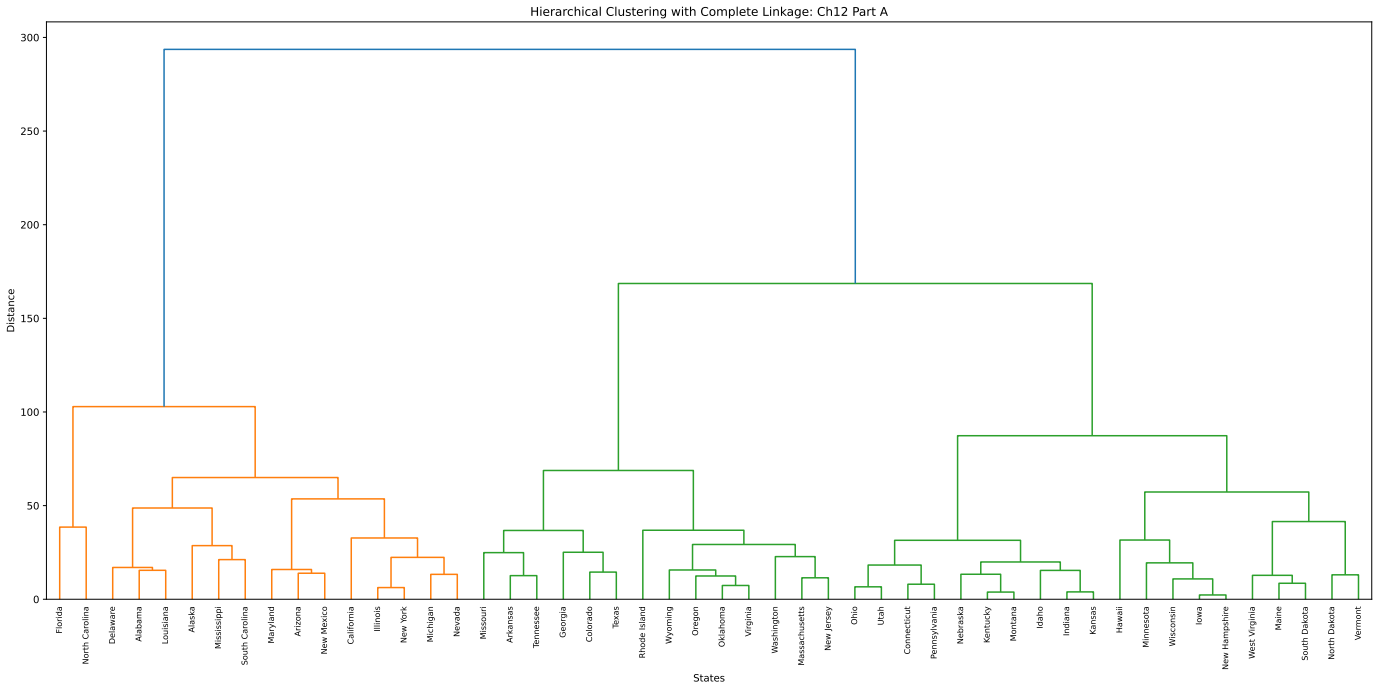


Figure 3: Clustering of States

b) The first of three clusters would have the states: Florida, North Carolina, Delaware, Louisianan, Alaska, Mississippi, South Carolina, Maryland, Arizona, New Mexico, California, Illinois, New York, Michigan and Nevada. The second cluster would have the states: Missouri, Arkansas, Tennessee, Georgia, Colorado, Texas, Rhode Island, Wyoming, Oregon, Oklahoma, Virginia, Washington, and New Jersey. The third cluster would have the states: Ohio, Utah, Connecticut, Pennsylvanian, Nebraska, Kentucky, Montana, Idaho, Indiana, Kansas, Hawaii, Minnesota, Wisconsin, Iowa, New Hampshire, West Virginia, Maine, South Dakota, North Dakota, and Vermont.

c) See inserted Figure. Done by computing z scores. 4

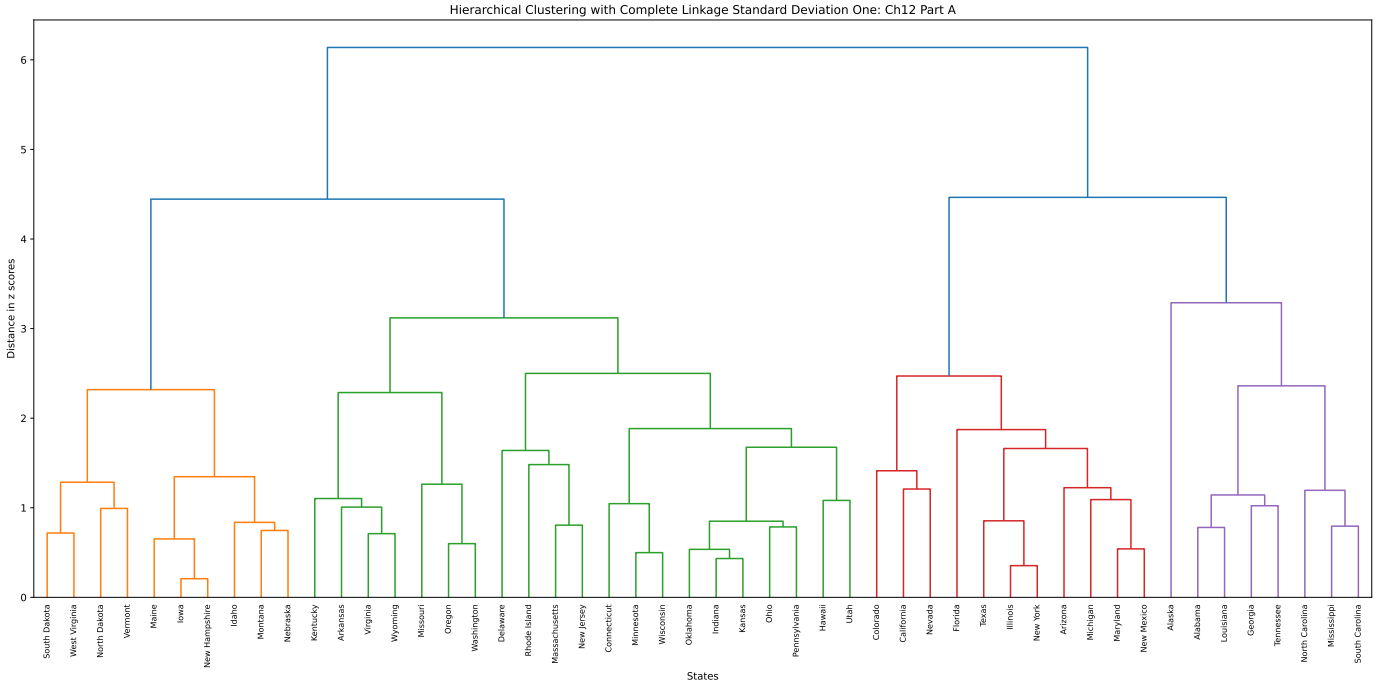


Figure 4: Clustering of States with Standard Deviation of 1

d) The main affect with changing the hierarchical clustering with complete linkage with a standard deviation of one is that there were an increase in the number of clusters. Due to the increase in the number of clusters the size of the clusters was a lot smaller (having less states). I created this graph by computing z scores for each state, satisfying the standard deviation of one requirement. As the clutters did not change to much, I think the variables should have been scaled a different method such as by median or mode.

Problem 8 (4 points)

Chapter 3, Exercise 4 (p. 122).

a) Referencing sketch of error value and flexibility in problem 2 part 'a' **??**. For training data, as flexibility increases residual sum of squares (RSS) decreases. A cubic model is more flexible than a linear model. So RSS is smaller for a cubic model than a linear model. For this problem, we would expect the RSS to be lower for a cubic model.

b) The cubic model would have a higher RSS than a linear model when using test data. The cubic model would over fit and lead to higher error.

c) The same answer from 'a' would apply. For training data, a more flexible model such as the cubic model would have a lower RSS than a less flexible model like the linear model even if the true relationship between X and Y is unknown.

d) Determining whether to use a cubic model or a linear model for this would be quite difficult as we do not know how far the relationship between X and Y is from being linear. If the relationship is very far from linear the cubic model would work better. If the relationship is quite close from linear a linear model would work better

Problem 9 (4 points)

Chapter 3, Exercise 9 (p. 123). In parts (e) and (f), you need only try a few interactions and transformations.

a) See inserted Figure 5. Note: Some of the names are slightly cut off due to image sizing difficulties.

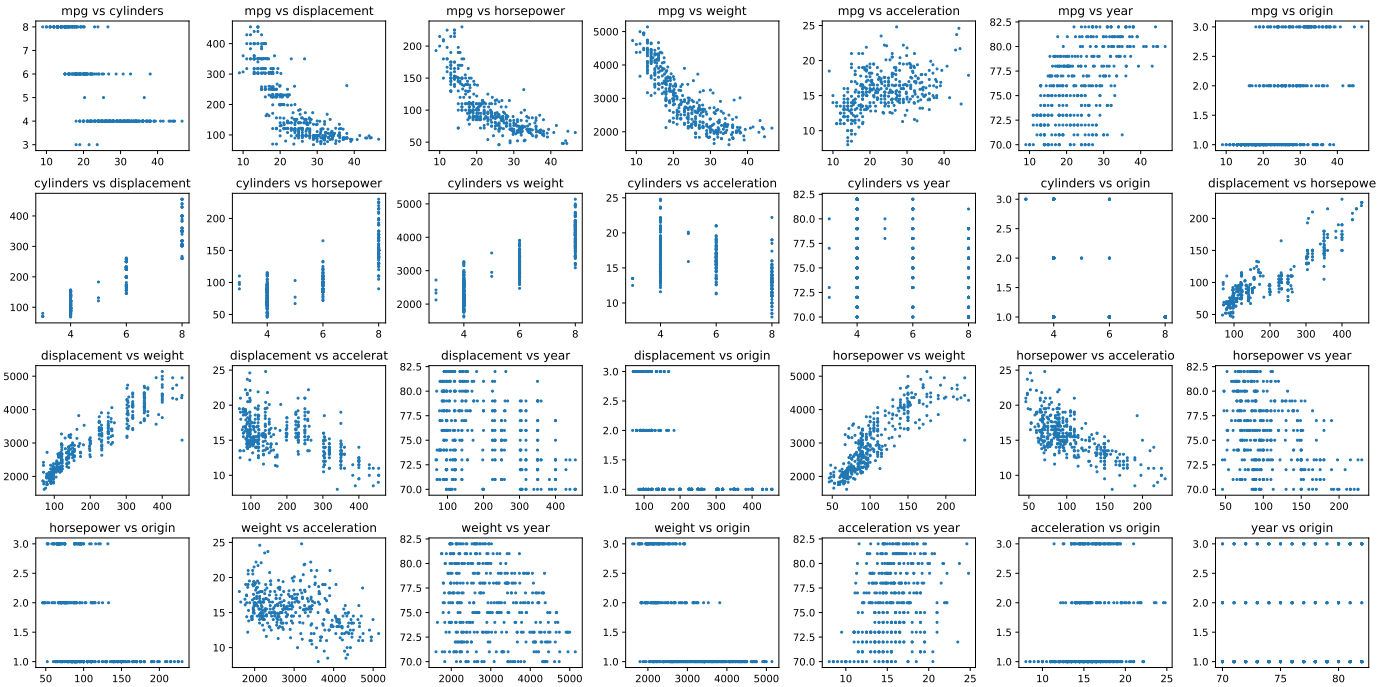


Figure 5: Multi Scatter plot

b) Here is the computed matrix of correlations between the variables using python function corrcoef cor_a = corrcoef(data_a, rowvar=False).

```
[ 1.     -0.7776 -0.8051 -0.7784 -0.8322  0.4233  0.5805  0.5652]
[-0.7776  1.      0.9508  0.843   0.8975 -0.5047 -0.3456 -0.5689]
[-0.8051  0.9508  1.      0.8973  0.933  -0.5438 -0.3699 -0.6145]
[-0.7784  0.843   0.8973  1.      0.8645 -0.6892 -0.4164 -0.4552]
[-0.8322  0.8975  0.933   0.8645  1.     -0.4168 -0.3091 -0.585 ]
[ 0.4233 -0.5047 -0.5438 -0.6892 -0.4168  1.      0.2903  0.2127]
[ 0.5805 -0.3456 -0.3699 -0.4164 -0.3091  0.2903  1.      0.1815]
[ 0.5652 -0.5689 -0.6145 -0.4552 -0.585   0.2127  0.1815  1.    ]]
```

---

c) The coefficients for the predictor variables for multiple linear regression are below. These can be used to fit the regression.

```
[('mpg', -0.49337631885848043), ('cylinders', 0.01989564374201724), ('displacement', -0.016951144227500457), (
('weight', 0.08057583832486492), ('acceleration', 0.7507726779503119), ('year', 1.4261404954231538), ('origin'
```

---

i) Cylinders, weight, year and origin have a relation ship with the response variable mpg because the p values are less than 0.05 (0.008, 0.000,0.000)

ii) The predictors with the most influence are below. These are in decreasing order of influence

pred_influence ['origin', 'year', 'acceleration', 'mpg', 'weight', 'cylinders', 'displacement', 'horsepower']

iii) The coefficient for year, 1.426, suggests that for every one year the mpg increases by 1.426.

d) Diagnostic plots of the linear regression fit 6. The residual plot does suggest that there could be potential outlines as several of the residuals are quite high.

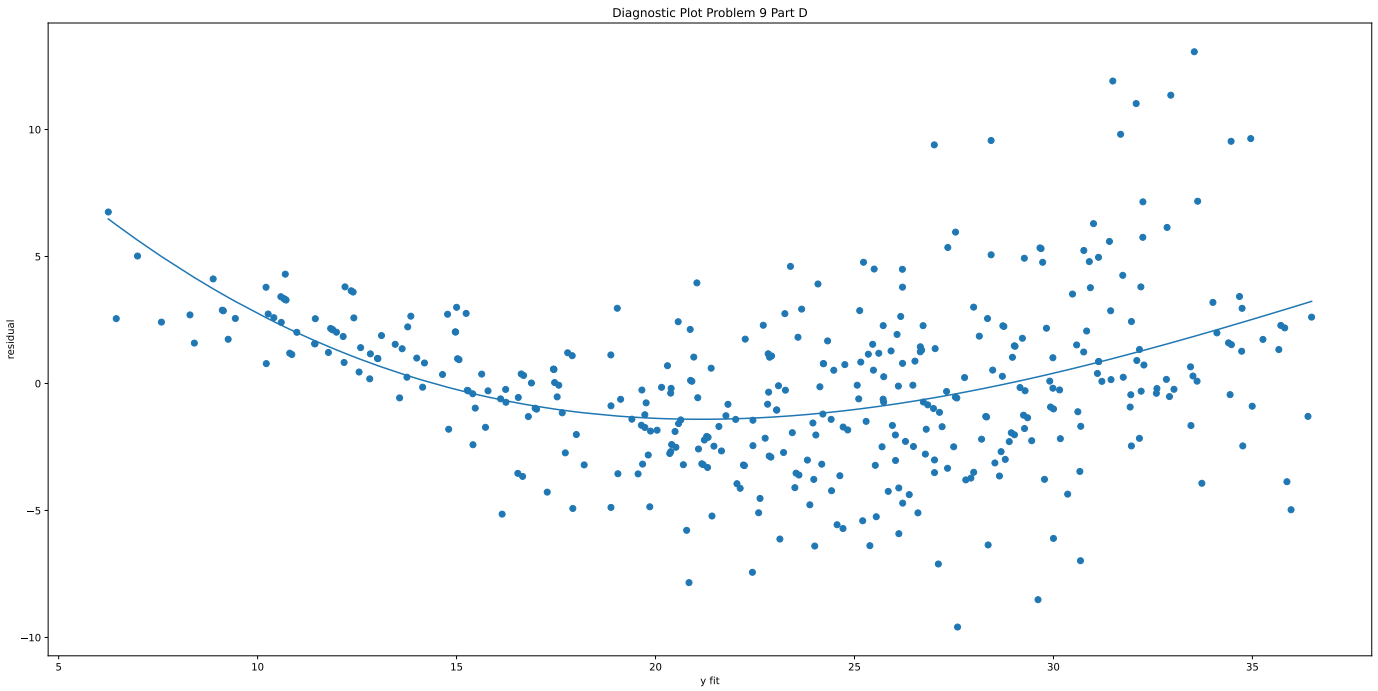


Figure 6: Residual Plot

e and f) I did several transformations and interaction affects. The data can be seen below. The transformations and interactions did appear to have significant affects as the r^2changed by about 5% and some of the coefficients changed by multiple magnitudes of 10.

Control: Coefficients b_v=array([-4.93e-01, 1.99e-02, -1.70e-02, -6.47e-03, 8.06e-02, 7.51e-01, 1.43e+00, -1.72e+01]) Control: r squared = 0.8214780764810599

Horse Power Squared: Coefficients b_v=array([1.01e-03, 3.49e-01, -7.56e-03, -3.19e-01, -3.27e-03, -3.31e-01, 7.35e-01, 1.01e+00, 1.32e+00]) Horse Power Squared: r squared = 0.8552261337659226

Horse Power Square Root: Coefficients b_v=array([-1.05e+01, 6.04e-02, -5.87e-03, 4.24e-01, -3.29e-03, -3.34e-01, 7.40e-01, 9.16e-01, 4.30e+01]) Horse Power Square Root: r squared = 0.8590511148607127

Horse Power Log: Coefficients b_v=array([-2.69e+01, -5.53e-02, -4.61e-03, 1.76e-01, -3.37e-03, -3.28e-01, 7.42e-01, 8.98e-01, 8.67e+01]) Horse Power Log: r squared = 0.8591868799104232

Horse Power * Weight: Coefficients b_v=array([5.53e-05, -2.96e-02, 5.95e-03, -2.31e-01, -1.12e-02, -9.02e-02, 7.69e-01, 8.34e-01, 2.88e+00]) Horse Power * Weight: r squared = 0.8618378127814719

Horse Power * Weight * Acceleration: Coefficients b_v=array([5.53e-05, -2.96e-02, 5.95e-03, -2.31e-01, -1.12e-02, -9.02e-02, 7.69e-01, 8.34e-01, 2.88e+00]) Horse Power * Weight * Acceleration: r squared = 0.8618378127814719

Problem 10 (4 points)

Chapter 3, Exercise 14 (p. 127).

a) Important note: I did all of this in Python, so the numbers and outcomes will be different than that of the number generation in R The Regression coefficients are below. The first term is constant, the second term is x1 and the third term is x2

[2.1892844 0.70462854 2.50240496]

b) The scatter plot of x1 and x2 is below. 7

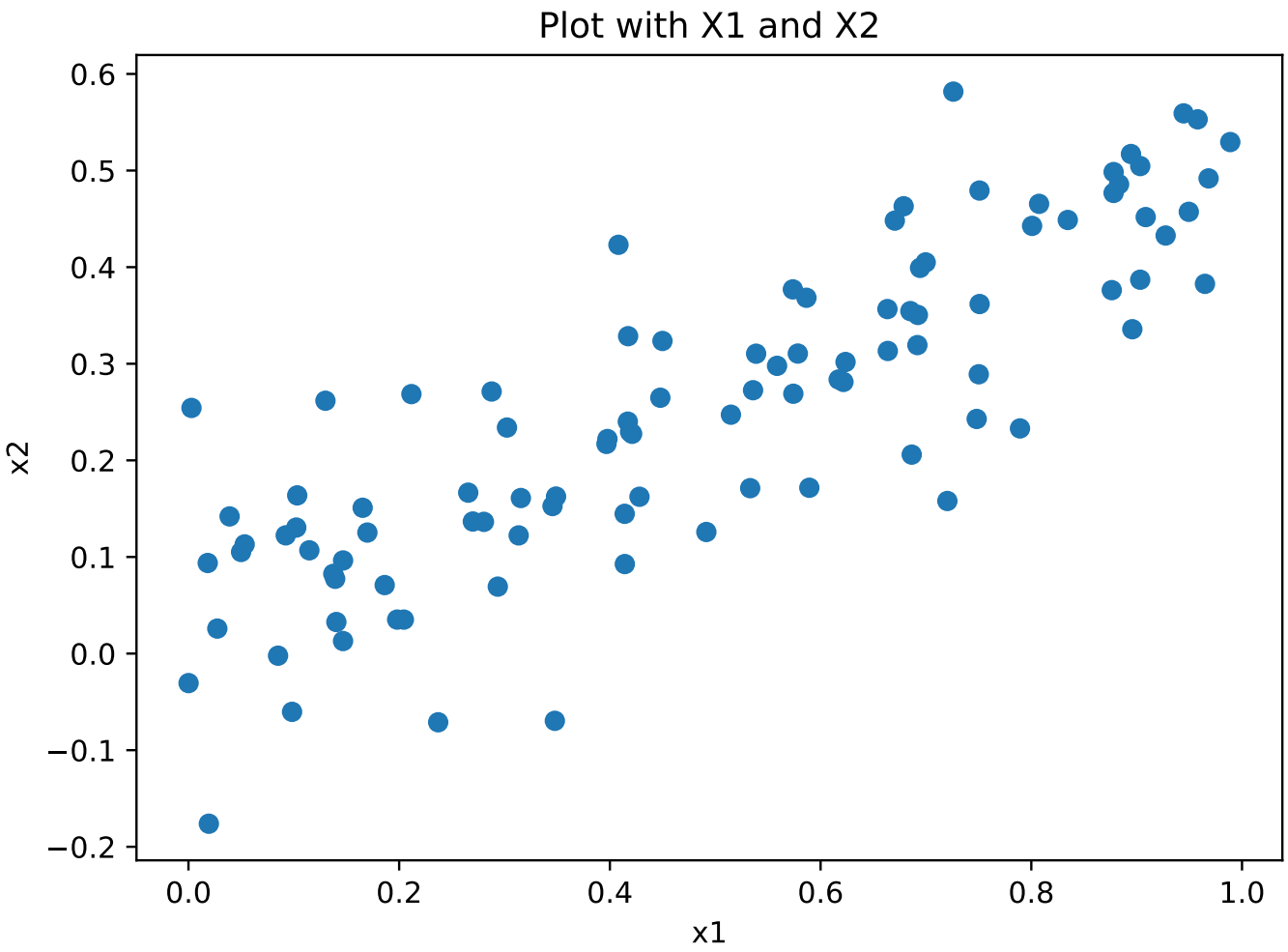


Figure 7: Plot of X1 and X2

c) We can reject the null hypothesis for $\beta_2 = 0$ because the p value (0.031) is less than 0.05

d) Control: Coefficients b_v=array([2.1892844 , 0.70462854, 2.50240496]) Control: r squared = 0.26050814407433387 X1 Only: Coefficients b_v=array([2.2485807 , 1.87698651]) X1 Only: r squared = 0.22380210449292925

We reject the null hypothesis because the p value is 0.001 which is less than 0.05.

e)X2 Only: Coefficients b_v=array([2.26552605, 3.56127637]) X2 Only: r squared = 0.25117295449150645

We reject the null hypothesis because the p value is 0.000 which is less than 0.05

f) The results some what contradict each other because we failed to reject the null hypothesis for $\beta 1$ but our multiple linear regression used both β_1 and $\beta 2$

e/g) New Control: Coefficients b_v=array([2.1892844 , 0.70462854, 2.50240496]) New Control: r squared = 0.26050814407433387

We reject the null hypothesis because the p value is 0.000 which is less than 0.05

New X1 Only: Coefficients b_v=array([2.3583255 , 1.72252265]) New X1 Only: r squared = 0.1810937124191967

We reject the null hypothesis because the p value is 0.000 which is less than 0.05

New X2 Only: Coefficients b_v=array([2.23312681, 3.72160649]) New X2 Only: r squared = 0.2877215955818587

We reject the null hypothesis because the p value is 0.000 which is less than 0.05

Problem 11 (5 points)

Mean Squared Error (MSE) is $\frac{1}{n}\Sigma(x-\mu)^2$. However this can be broken down into three parts which are square bias, variance and irreducible error. The breaking down of mean squared error into these three parts is known as the bias variance decomposition which is closely related to the bias variance trade off. Bias and variance change depending on the flexibility, the number of parameters of the fitted function. The reason why there is a trade off is flexibility. Low variance scenarios are those of a simple function, one with low flexibility. Low bias scenarios are those of high complexity or high flexibility. Because of this, there is a trade off between bias and variance if one would want to keep both low. Irreducible error is error that is always there. These three components make up mean square error. The three components show what bias decomposition is.

Problem 12 (5 points)

Let x1 , . . . , xn be a fixed set of input points and yi = f (xi) + εi , where εi iid$\sim$ Pε with E (εi) = 0 and Var (εi) < ∞. Prove that the MSE of a regression estimate f$\hat{}$ fit to (x1 , y1), . . . , (xn , yn) for a random test

x0 or E y0 – f$\hat{}$(x0) decomposes into variance, square bias, and irreducible error components. Hint: You can apply the bias-variance decomposition proved in class.

Problem 12 (5 points)

a

Assume that y_i and $x_i$are mean subtracted, i.e. $y_i^{'}$and $x_i^{'}$were the original variables and $y_i = y_i^{'} - \bar{y^{'}}$ and $x_i = x_i^{'} - \bar{x^{'}}$ respectively

MSE $= \frac{1}{n}\Sigma(y_i - \hat{y_i})^2$

$= \frac{1}{n}\Sigma(y_i - \hat{f}(x_i))^2$

$= \frac{1}{n}\Sigma(y_i - \beta x_i)^2$

$\frac{dMSE}{d\beta} = \frac{1}{n}\Sigma 2(y_i - \beta x_i)(-x_i)$

$\frac{1}{n}\Sigma(y_i - \beta x_i)x_i = 0$

2 and '-' sign go away because right hand side is zero

$\frac{1}{n}\Sigma(y_i x_i - \beta x_i x_i) = 0$

$\frac{1}{n}\Sigma y_i x_i = \frac{1}{n}\Sigma\beta x_i x_i$

$\beta = \frac{\sigma_{xy}}{\sigma_x^2}$

Because $\frac{1}{n}\Sigma y_i x_i = \sigma_{xy}$ and $\frac{1}{n}\Sigma x_i x_i = \sigma_x^2$when the variables have been mean subtracted

(b) $\sigma^2/(x_i - \bar{x})^2$