DAT 500S Module 2

Textbook Quiz:
Exercise 2.4 Q3&Q7 Exercise 3.7; Problem 3a and 3b; Problem 4.

Group Assignment:

Exercise 3.7 Q9

Individual Assignment:

Exercise 3.7 Q10

Module 2 – Important Concept

- It can be shown that for any given, X=x_0, the expected test MSE for a new Y at x_0 will be equal to

$$Expected\,Test\,MSE = E\left(y_0 - \hat{f}(x_0)\right)^2 = Bias^2 + Var + \underbrace{\sigma^2}_{\text{Irreducible Error}}$$

- What this means is that as a method gets more complex the bias will decrease and the variance will increase but expected test MSE may go up or down!

Module 2 – Important Concept

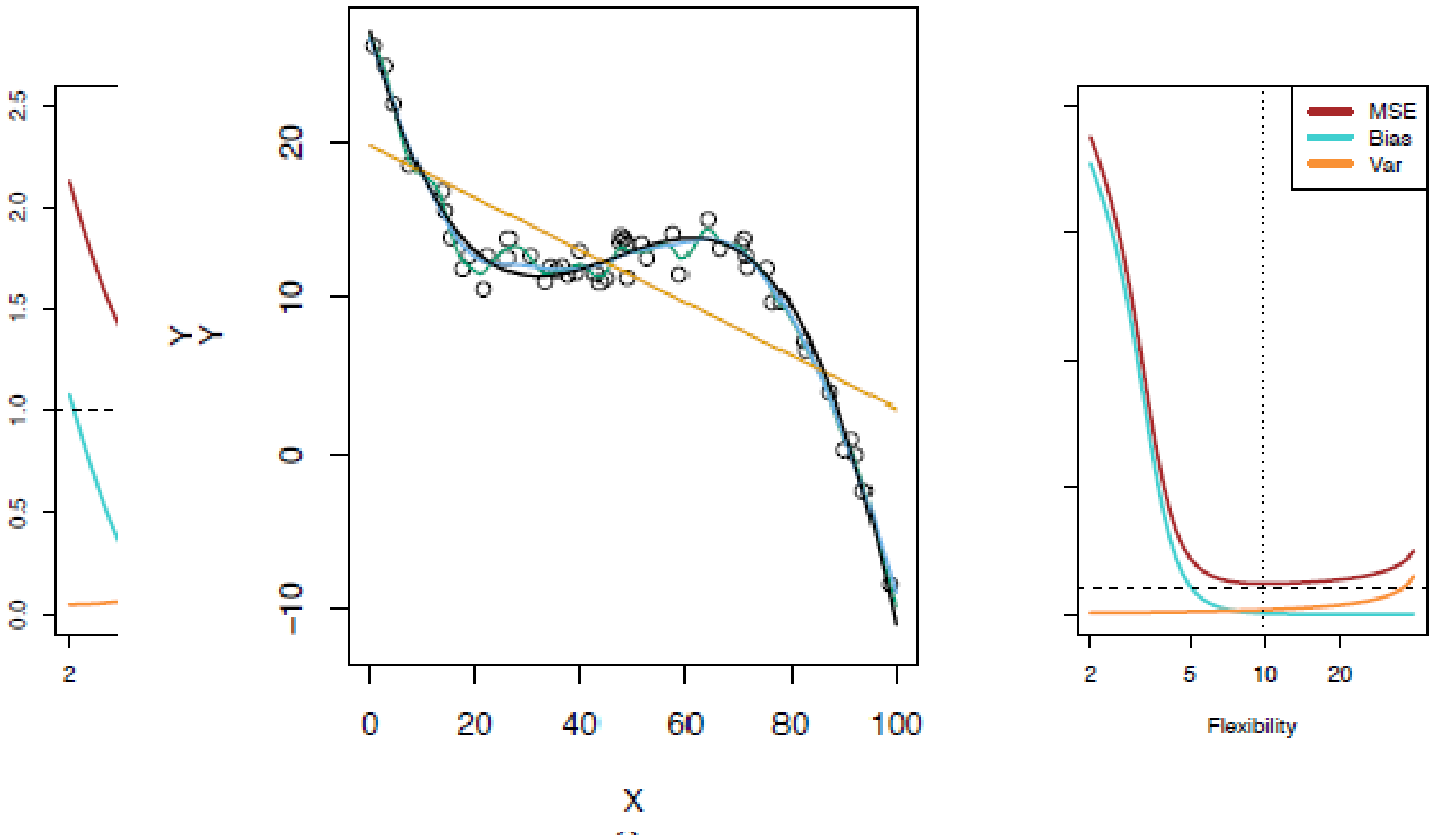
$$Expected\ Test\ MSE = E\left(y_0 - \hat{f}(x_0)\right)^2$$

$$= [Bias(\hat{f}(x_0))]^2 + Var(\hat{f}(x_0)) + \underbrace{\sigma^2}_{\text{Irreducible Error}}$$

$$[Bias(\hat{f}(x_0))]^2 = \left[E\left(\hat{f}\left(x_0\right)\right) - f\left(x_0\right)\right]^2$$

$$Var(\hat{f}(x_0)) = E\left[\hat{f}\left(x_0\right) - E\left(\hat{f}\left(x_0\right)\right)\right]^2$$

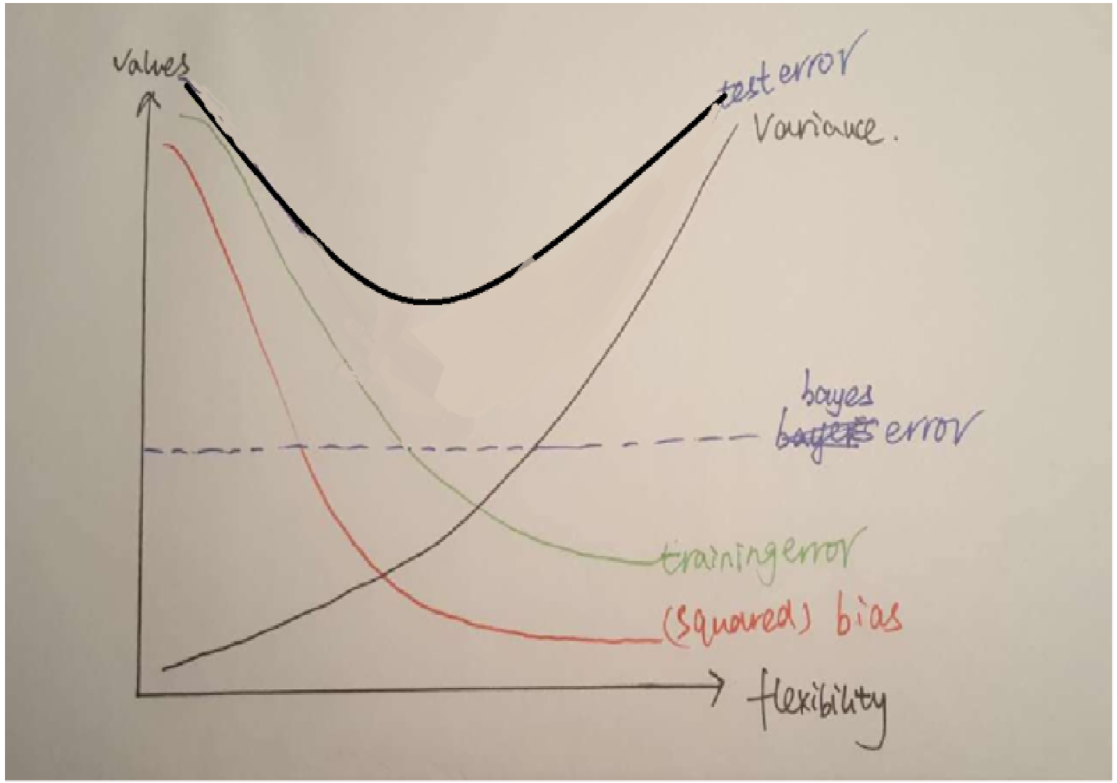
Module 2 – Important Concept



Quiz: Exercise 2.4 Q3

a)Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot.

b) Explain why each of the five curves has the shape displayed?



- Flexibility increase:
  ⇒more variance but less bias
- irreducible error: constant, depends only on data itself
- testing error: u-shape, overfitting with too much flexibility
- training error: decreasing

Quiz: Exercise 2.4 Q3

b) Explain why each of the five curves has the shape displayed? (detailed)

The Bayes or irreducible error is constant; this is a property of the data and does not depend on the learning method. The test MSE is the sum of variance, square bias, and Bayes error, per the bias-variance decomposition. The variance component increases with flexibility and the squared bias decreases; the shapes of these curves will depend on the learning method and the data. The training MSE will decrease, potentially below the Bayes error, as the flexibility of the method increases. In this plot, we have shown this curve go down to zero, which would suggest we are using a nonparametric method which is able to fit any function.

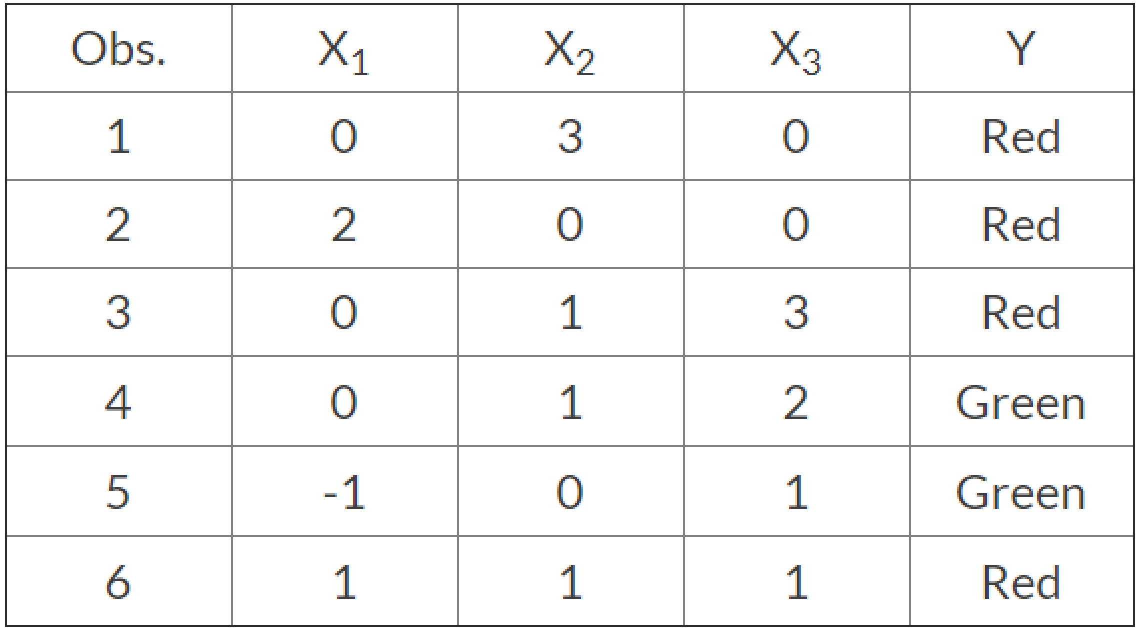
Quiz: Exercise 2.4 Q7

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2= X3 = 0.

distance:

sqrt [(X1-0)^2+(X2-0)^2+(X3-0)^2]

- obs 1. ⇒3
- obs 2. ⇒2
- obs 3. ⇒sqrt(10)
- obs 4. ⇒sqrt(5)
- obs 5. ⇒sqrt(2)
- obs 6. ⇒sqrt(3)



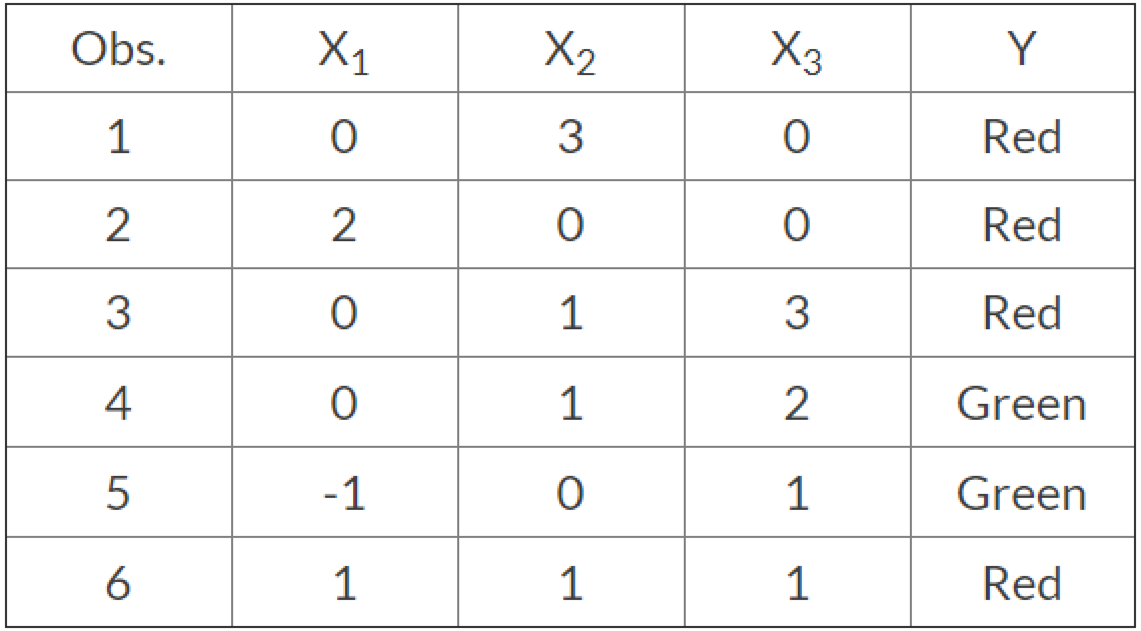
Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Quiz: Exercise 2.4 Q7

- For any given X we find the k closest neighbors to X in the training data, and examine their corresponding Y.
- The smaller that k is the more flexible the method will be.

Quiz: Exercise 2.4 Q7

- obs 1. ⇒ 3
- obs 2. ⇒ 2
- obs 3. ⇒ sqrt(10)
- obs 4. ⇒ sqrt(5)
- obs 5. ⇒ sqrt(2)
- obs 6. ⇒ sqrt(3)



Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

(b) prediction with K = 1?

Green. Observation #5 is the closest neighbor for K = 1.

(c) prediction with K = 3?

Red. Observations #2, 5, 6 are the closest neighbors, majority is Red

Quiz: Exercise 2.4 Q7

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

the method will be more flexible as K becomes smaller

Quiz: Exercise 3.7 Q3

3. Suppose we have a data set with five predictors, X1 =GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and : $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10.$

ŷ = 50 + 20*GPA + 0.07*IQ + 35*Gender
+ 0.01*GPA*IQ- 10*GPA*Gender

Quiz: Exercise 3.7 Q3

$\hat{y}$ male (Gender= 0) $= 50 + 20*GPA + 0.07*IQ + 0.01*GPA * IQ$
$\hat{y}$ female(Gender=1)=
$50 + 20*GPA + 0.07*IQ + 35+0.01*GPA * IQ - 10 *GPA$

i. For a fixed value of IQ and GPA, males earn more on average than females. False. 35-10*GPA, we do not know if it is positive

ii. For a fixed value of IQ and GPA, females earn more on average than males.False, same reason as i

Quiz: Exercise 3.7 Q3

$\hat{y}$ male (Gender= 0) = $50 + 20*GPA + 0.07*IQ + 0.01*GPA * IQ$

$\hat{y}$ female(Gender=1)=

$50 + 20*GPA + 0.07*IQ + 35+0.01*GPA * IQ - 10 *GPA$

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.True, if GPA is high enough, 35-10*GPA will be negative

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.False

Quiz: Exercise 3.7 Q3

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

$\hat{y}$= 50 + 20 * 4 + 0.07 * 110 + 35 + 0.01 (4 * 110) - 10 * 4
= 137.1

Quiz: Exercise 3.7 Q4

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression.When true relationship between X and Y is linear, Consider the training/testing residual sum of squares (RSS) for the linear regression, and also the training/testing RSS for the cubic regression. would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(a) the cubic regression will have a lower training RSS because it is more flexible and will fit more data

(b) the cubic regression will have a higher testing RSS ; overfitting, leading to high variance in the estimate of the function. However, if n is large, the two estimates could be close, making their test RSS similar

Quiz: Exercise 3.7 Q4

Suppose that the true relationship between X and Y is not linear, but we don’t know how far it is from linear. Consider the training/testing RSS for the linear regression, and also the training/testing RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(c) Cubic regression has lower train RSS than the linear fit because of higher flexibility

(d) There is not enough information to tell. we don’t know"how far it is from linear". If the true relationship between X and Y is cubic, then we would expect the test RSS to be lower for the cubic regression model. However, if the true relationship is quadratic, for example, the test RSS could be lower for the linear model.

Group Assignment: Problem 9

```
Auto = read.csv("Auto.csv",header = T, na.strings = "?")
fix(Auto)
Auto = na.omit(Auto)
```

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

⇒use pairs(Auto)

(b) correlation matrix:

exclude the name variable, which is qualitative.

⇒round(cor(Auto[1:8]) ,digits=3)

Group Assignment: Problem 9

(c) perform a multiple linear regression with mpg as the response and all other variables except name as the predictors.

“origin” is categorical data but has been coded as quan data in R

factorize “origin” before regression: Auto$Origin = factor(Auto$Origin)

```
autoLinearModel = lm(mpg~ . -name, data=Auto)
summary(autoLinearModel)
```

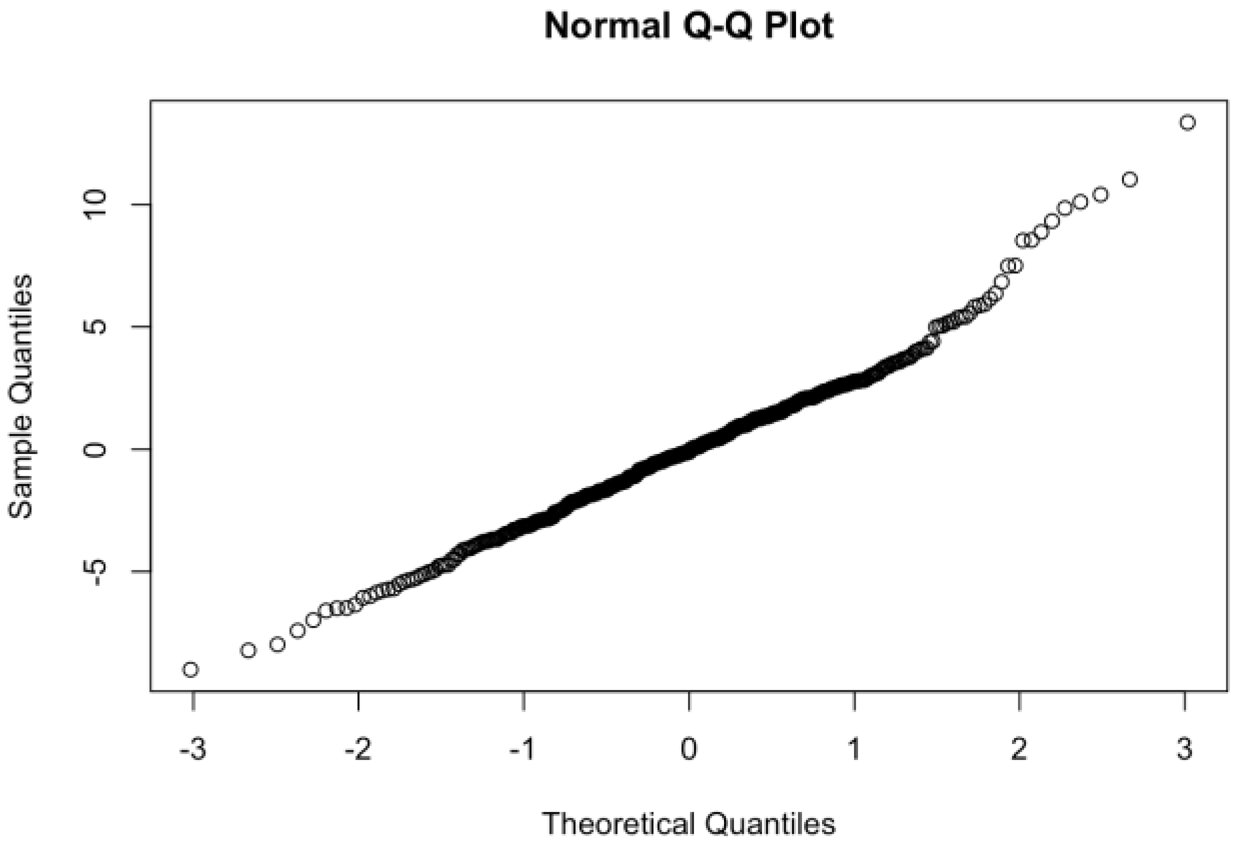
Comment:

- look at R square
- variables that are significant (year, origin, weight,displacement)
- all else equal, gas efficiency improves over time.

Group Assignment: Problem 9

(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Do the residual plots suggest any unusually large outliers? Does the leverage unusually high leverage?

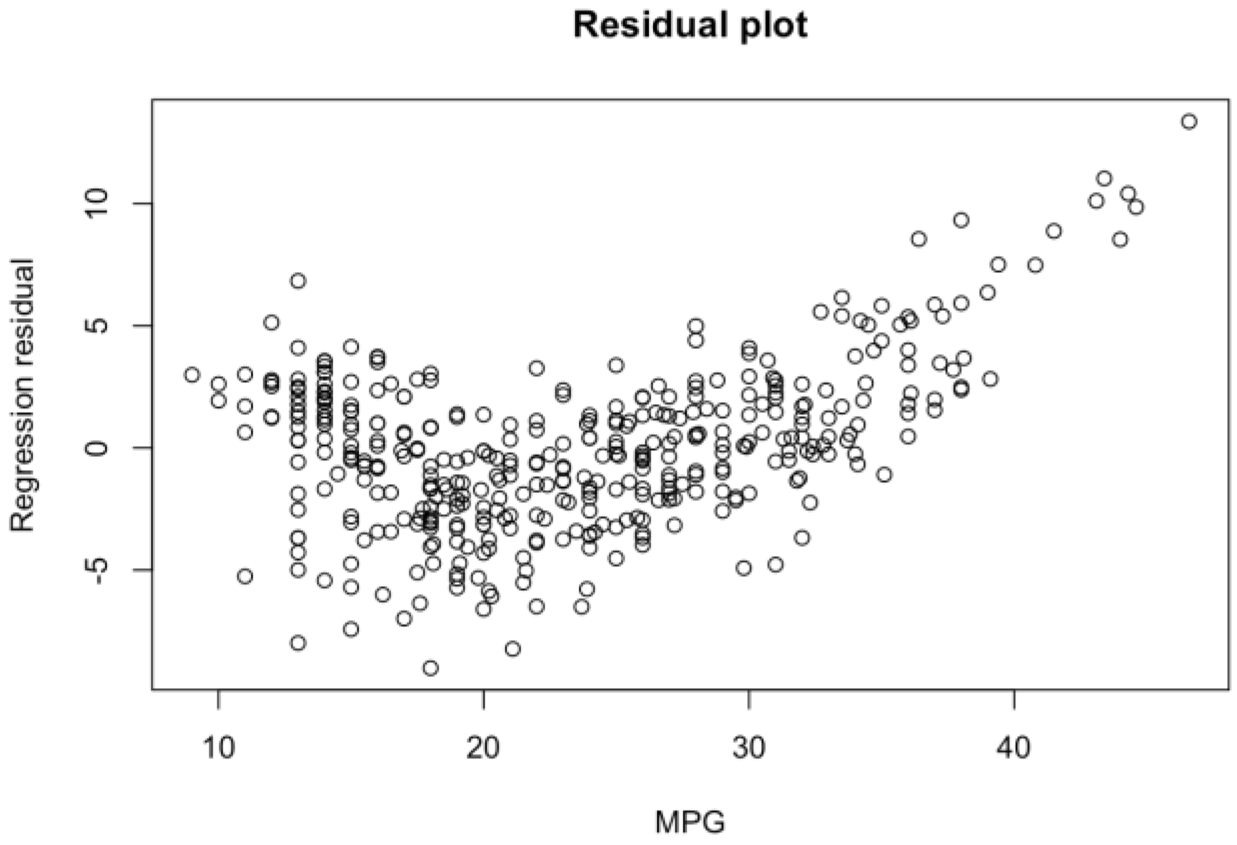
normal qq plot
=>qqnorm(autoLinearMode



Group Assignment: Problem 9

(d) residual plot

⇒plot(Auto[, 1], autoLinearModel$residuals, xlab = "MPG", ylab = "Regression residual", main = "Residual plot")

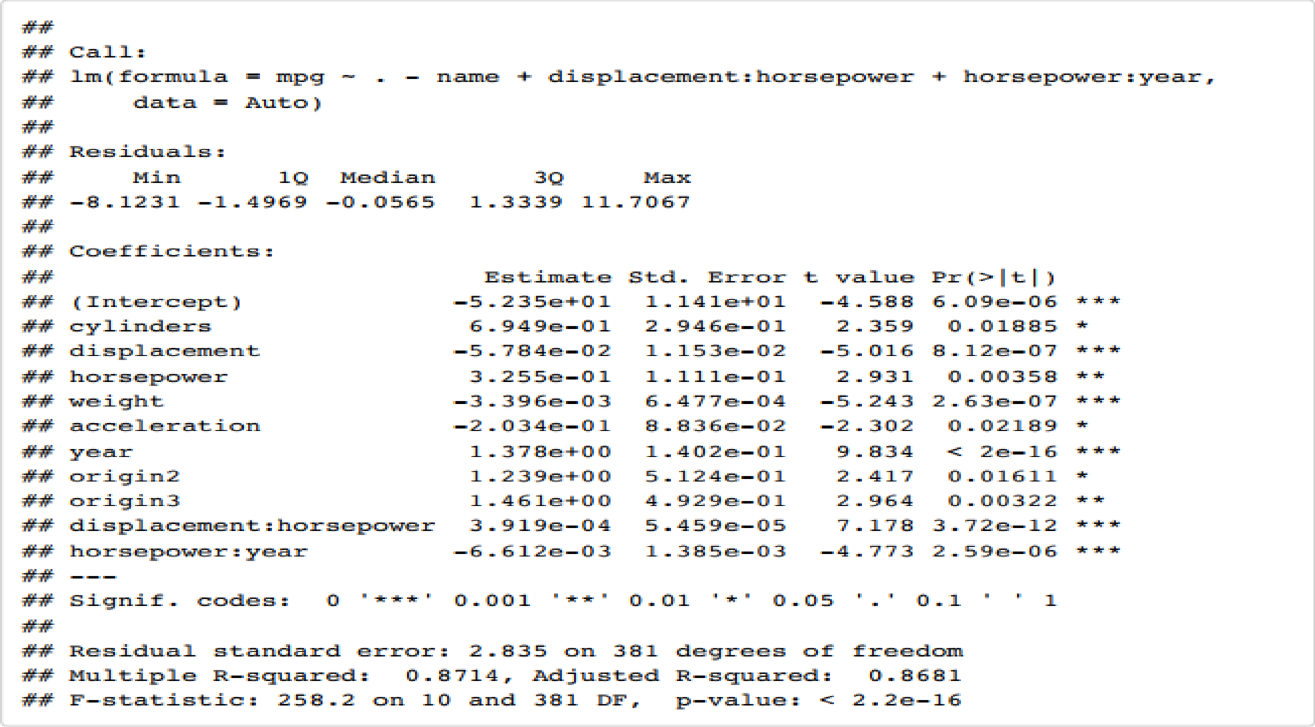


Group Assignment: Problem 9

(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

Consider significant interactions:

autoInteractionModel = lm(mpg ~ . - name + displacement:horsepower + horsepower:year, data = Auto) summary(autoInteractionModel)



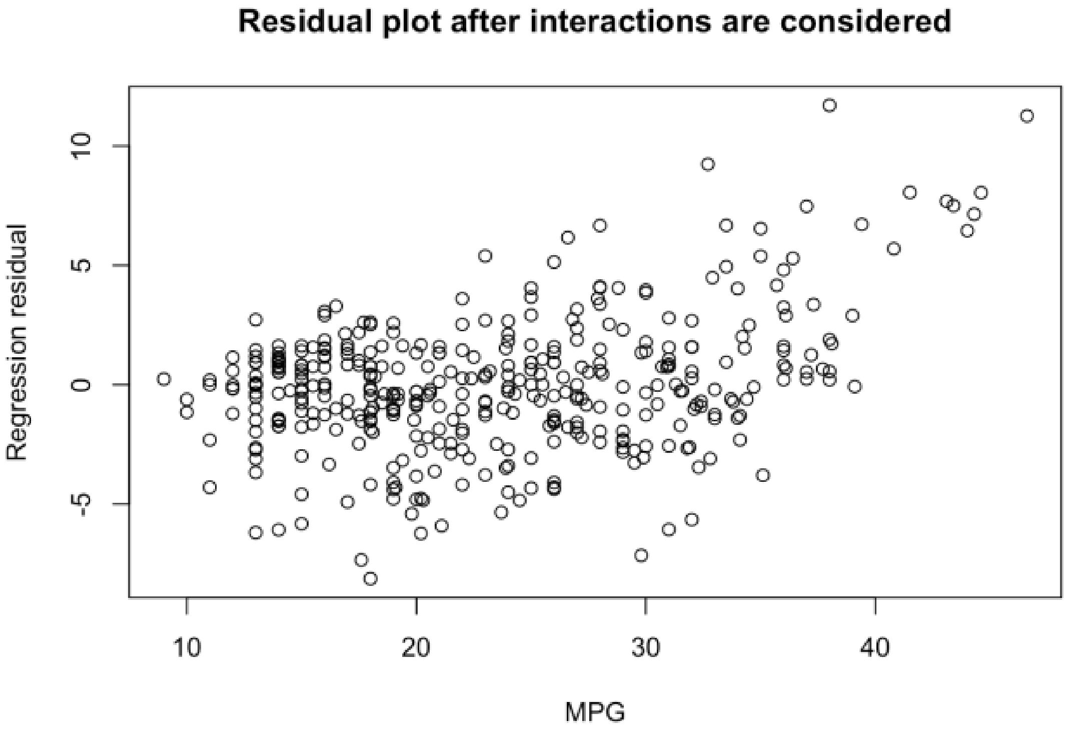
```
##
## Call:
## lm(formula = mpg ~ . - name + displacement:horsepower + horsepower:year,
##     data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1231 -1.4969 -0.0565  1.3339 11.7067
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -5.235e+01  1.141e+01  -4.588 6.09e-06 ***
## cylinders               6.949e-01  2.946e-01   2.359  0.01885 *
## displacement           -5.784e-02  1.153e-02  -5.016 8.12e-07 ***
## horsepower              3.255e-01  1.111e-01   2.931  0.00358 **
## weight                 -3.396e-03  6.477e-04  -5.243 2.63e-07 ***
## acceleration           -2.034e-01  8.836e-02  -2.302  0.02189 *
## year                    1.378e+00  1.402e-01   9.834  < 2e-16 ***
## origin2                 1.239e+00  5.124e-01   2.417  0.01611 *
## origin3                 1.461e+00  4.929e-01   2.964  0.00322 **
## displacement:horsepower  3.919e-04  5.459e-05   7.178 3.72e-12 ***
## horsepower:year        -6.612e-03  1.385e-03  -4.773 2.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.835 on 381 degrees of freedom
## Multiple R-squared:  0.8714, Adjusted R-squared:  0.8681
## F-statistic: 258.2 on 10 and 381 DF,  p-value: < 2.2e-16
```

Group Assignment: Problem 9

(e) recheck residual plot:

```
plot(Auto[, 1], autoInteractionModel$residuals, xlab = "MPG", ylab = "Regression residual", main = "Residual plot after interactions are considered")
```

=>interaction didn't fix the problem



Group Assignment: Problem 9

(f) Try a few different transformations of the variables
from the scatter plot matrix in part (a), it seems there is a quadratic relationship between MPG and displacement, and MPG and weight

a possible transformation:

```
autoLinearModel = lm(mpg ~ . - name + I(weight^2) +
I(displacement^2), data = Auto) summary(autoLinearModel)
```