HW1-Jiayuan Guo

1. Problem 4, Chapter 3:

(a)For training dataset, the cubic regression has a lower RSS than linear regression, because cubic regression fits the training dataset much tighter.

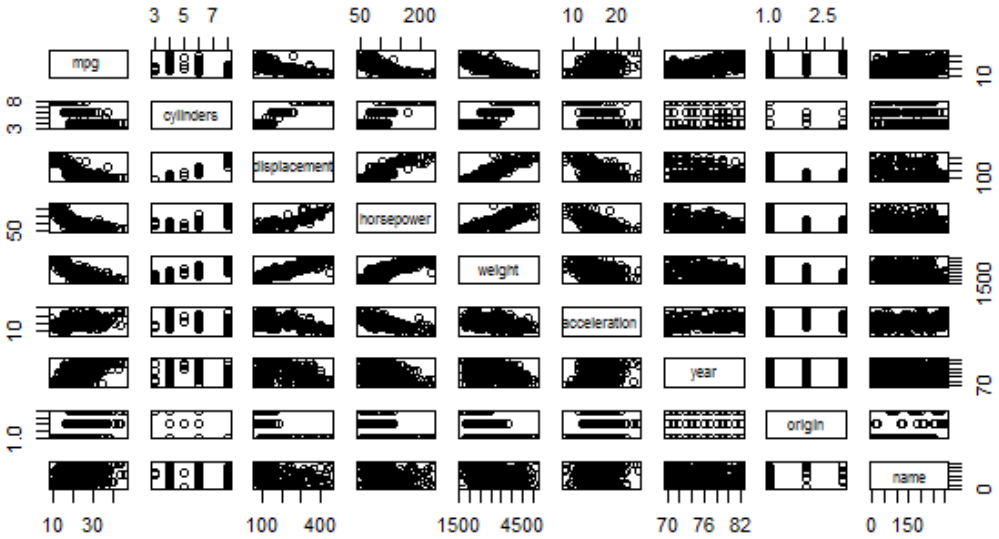
(b)For testing dataset, the cubic regression has a higher RSS than linear regression, because the true relationship between X and Y is linear and cubic regression has more variables in model, which leads to overfitting.

(c) The cubic regression still has a lower RSS than linear regression. No matter what exact relationship is between X and Y, cubic regression always has a better fit for training dataset than linear regression.

(d)There is not enough information to tell whether linear regression or cubic regression has a lower training RSS. There is a bias-variance tradeoff because we are not sure the real relationship between X and Y is closer to which model. If linear regression model is closer to real relationship, linear regression will have a lower testing RSS; if cubic regression model is closer to real relationship, cubic regression will have a lower testing RSS.

2. Problem 9, Chapter 3

(a)



(b)

	mpg	cylinders	displacement	horsepower	weight	acceleration
mpg	1.0000000	-0.7776175	-0.8051269	-0.7784268	-0.8322442	0.4233285
cylinders	-0.7776175	1.0000000	0.9508233	0.8429834	0.8975273	-0.5046834
displacement	-0.8051269	0.9508233	1.0000000	0.8972570	0.9329944	-0.5438005
horsepower	-0.7784268	0.8429834	0.8972570	1.0000000	0.8645377	-0.6891955
weight	-0.8322442	0.8975273	0.9329944	0.8645377	1.0000000	-0.4168392

```
acceleration    0.4233285  -0.5046834      -0.5438005  -0.6891955  -0.4168392        1.0000000
year                0.5805410  -0.3456474      -0.3698552  -0.4163615  -0.3091199        0.2903161
origin              0.5652088  -0.5689316      -0.6145351  -0.4551715  -0.5850054        0.2127458
                                year          origin
mpg                 0.5805410    0.5652088
cylinders       -0.3456474  -0.5689316
displacement  -0.3698552  -0.6145351
horsepower      -0.4163615  -0.4551715
weight              -0.3091199  -0.5850054
acceleration    0.2903161    0.2127458
year                1.0000000    0.1815277
origin              0.1815277    1.0000000
```

(c)

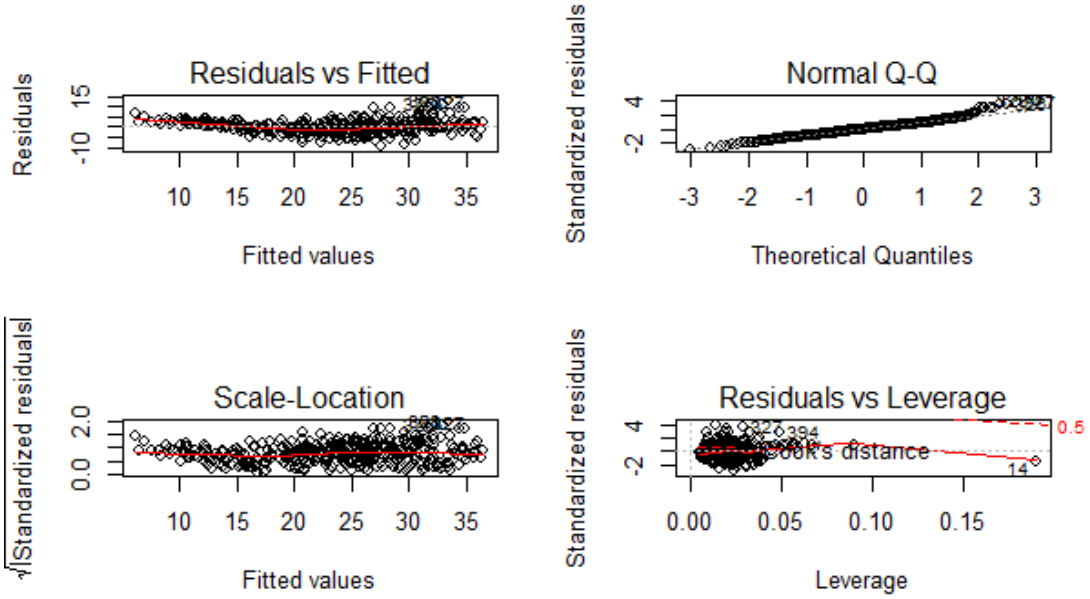
```
Call:
lm(formula = mpg ~ . - name, data = Auto)
Residuals:
      Min          1Q    Median          3Q         Max
-9.5903  -2.1565  -0.1169    1.8690  13.0604
Coefficients:
                     Estimate  Std.  Error  t  value  Pr(>|t|)
(Intercept)    -17.218435      4.644294    -3.707    0.00024  ***
cylinders          -0.493376      0.323282    -1.526    0.12780
displacement      0.019896      0.007515      2.647    0.00844  **
horsepower       -0.016951      0.013787    -1.230    0.21963
weight              -0.006474      0.000652    -9.929    <  2e-16  ***
acceleration      0.080576      0.098845      0.815    0.41548
year                 0.750773      0.050973    14.729    <  2e-16  ***
origin               1.426141      0.278136      5.127  4.67e-07  ***
---
Signif.  codes:    0  ?**?0.001  ?*?0.01  ??0.05  ??0.1  ??1
Residual  standard  error:  3.328  on  384  degrees  of  freedom
Multiple  R-squared:    0.8215,   Adjusted  R-squared:    0.8182
F-statistic:  252.4  on  7  and  384  DF,    p-value:  <  2.2e-16
```

i. By testing null hypothesis, it shows that F-statistic is far away from 1 and p- value is small. So the null hypothesis is invalid and there is a relatioship between the predictors and the response

ii. The displacement, weight, year, origin have p-values which is smaller than 0.025, so these four predictors appears to have a statistically significant relationship to the response (95% CI)

iii. The coefficient for the year variable is 0.750773, which means when other predictors remain to be constant, an increase of 1 year followed with an increase of 0.7507727 in “mpg”.

(d)



The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2). Point 14 has an unusual high leverage.

(e)

Call:

lm(formula = mpg ~ cylinders * displacement + displacement * weight, data = Auto[, 1:8])

Residuals:

Min	1Q	Median	3Q	Max
-13.2934	-2.5184	-0.3476	1.8399	17.7723

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.262e+01	2.237e+00	23.519	< 2e-16	***
cylinders	7.606e-01	7.669e-01	0.992	0.322	
displacement	-7.351e-02	1.669e-02	-4.403	1.38e-05	***
weight	-9.888e-03	1.329e-03	-7.438	6.69e-13	***
cylinders: displacement	-2.986e-03	3.426e-03	-0.872	0.384	
displacement: weight	2.128e-05	5.002e-06	4.254	2.64e-05	***

Signif. codes: 0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

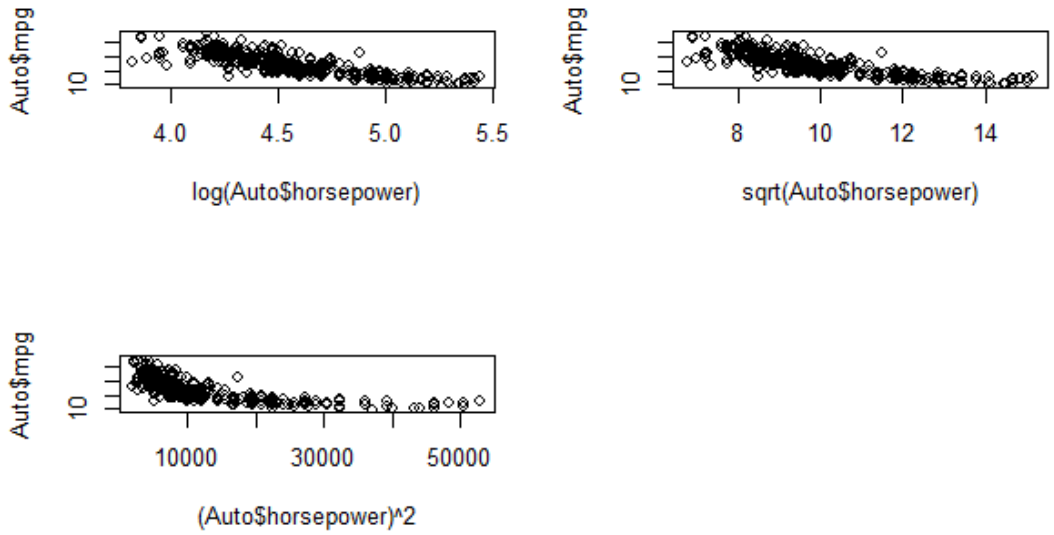
Residual standard error: 4.103 on 386 degrees of freedom

Multiple R-squared: 0.7272, Adjusted R-squared: 0.7237

F-statistic: 205.8 on 5 and 386 DF, p-value: < 2.2e-16

The p-values is smaller than 2.2e-16, which means the interaction between displacement and weight is statistically significant.

(f)



The outputs of log transform of mpg seem to have better model fitting.

3. Problem 3, Chapter 5

(a)Divide dataset into k bins of same size, then take one bin as a testing bin and the other (k-1) bins as training bins. After these k times learning process, average these k times performance, so you can get a more accurate testing result.

(b)

i. The validation set approach is conceptually simple and is easy to implement.

(1.) the validation estimate of the test error rate can be highly variable, depending on which observations are included in the training and validation sets.

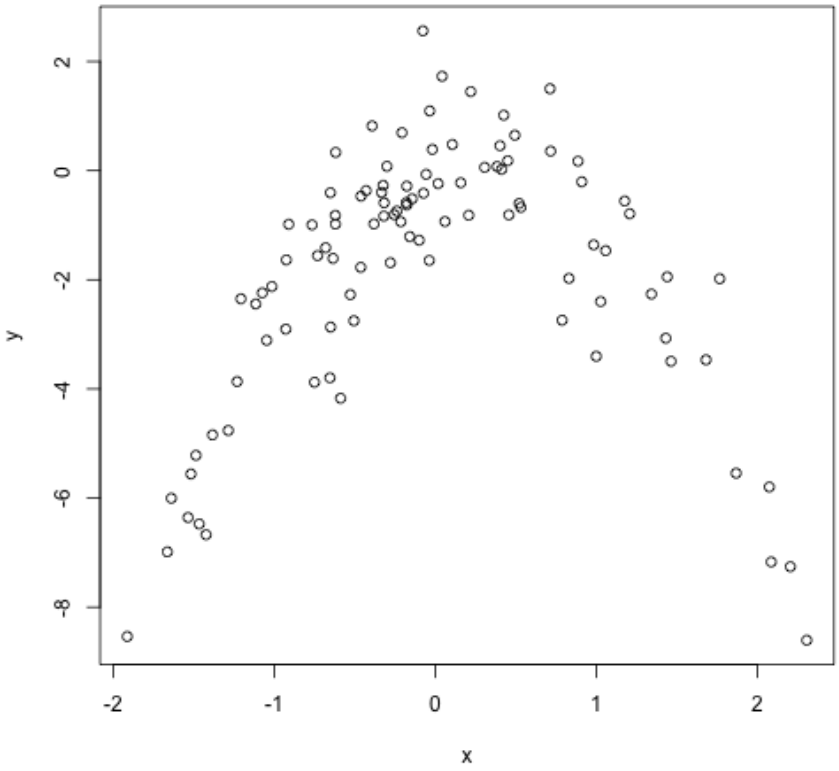
(2) Because only subset of observations are used to fit the model, so the validation set error rate may tend to overestimate the test error rate for the model fit on the entire dataset.

ii. LOOCV is a special case of cross-validation that choose a single observation for validation and the other for training, so it has far less bias and tends not to overestimate the test error rate as much as the validation set approach does. But LOOCV needs to run process for n times, which causes heavy computational load and high variance.

4. Problem 8, Chapter 5

(a)In this model, n=100 and p=2, model equation: $y=x-2x^2+\varepsilon$.

(b)



It is a quadratic plot.

(c)

i. [1] 5.890979 5.888812

ii. [1] 1.086596 1.086326

iii. [1] 1.102585 1.102227

iv. [1] 1.114772 1.114334

(d)When I alter seed(1) to seed(100), the result is exact same, because LOOCV will be the same because LOOCV evaluates n folds of a single observation.

(e) The LOOCV estimate for quadratic polynomial has the lowest test error rate. This is exactly what we expect, because in the plot of (b), we could assume the relationship between x and y is quadratic.

(f) The p-values of the linear and quadratic terms are small, which indicate the statistical significance of these two model. This conclusion agrees strongly with our cross-validation results.