CSC 423 Project 2

Part A:

1.

Weight NBags

1 5050 100

2 10249 205

3 20000 450

4 7420 150

5 24685 500

6 10206 200

7 7325 150

8 4958 100

9 7162 150

10 24000 500

11 4900 100

12 14501 300

13 28000 600

14 17002 400

15 16100 400

2.

Mean of Weight: 13437.2

Mean of NBags: 287

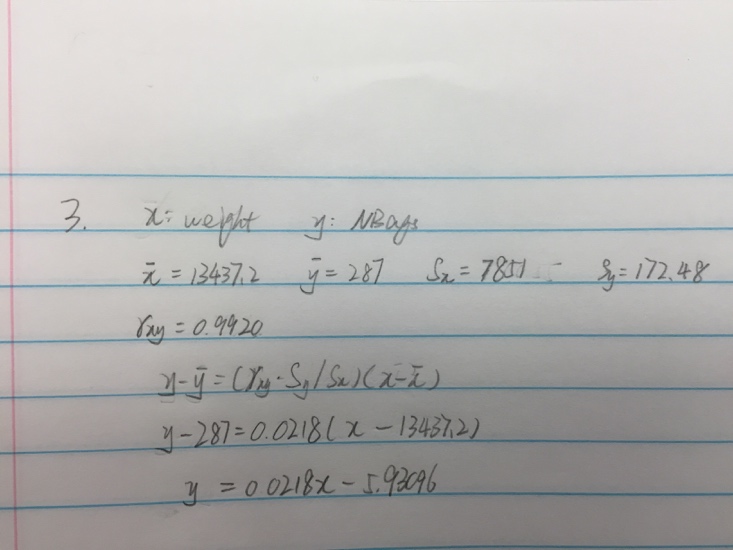
SD of Weight: 7850.551

SD of NBags: 172.4798

Covariance of Weight and NBags: 1343254

Correlation of Weight and NBags: 0.9920181

3.



4.

lm(formula = NBags ~ Weight, data = flour)

Residuals:

Min 1Q Median 3Q Max

-32.146 -11.349 -4.201 -0.582 54.964

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.8643573 11.8557460 -0.495 0.629

Weight 0.0217950 0.0007684 28.366 4.45e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.57 on 13 degrees of freedom

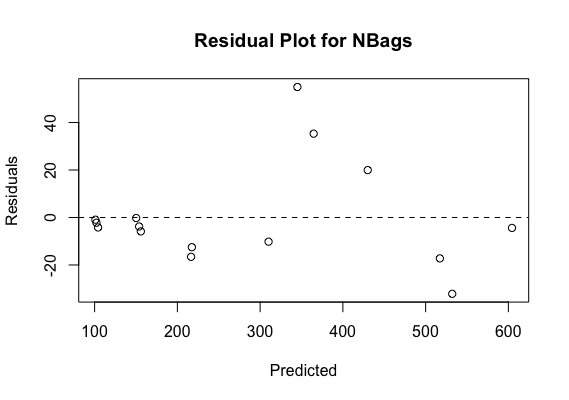
Multiple R-squared: 0.9841, Adjusted R-squared: 0.9829

F-statistic: 804.6 on 1 and 13 DF, p-value: 4.454e-13

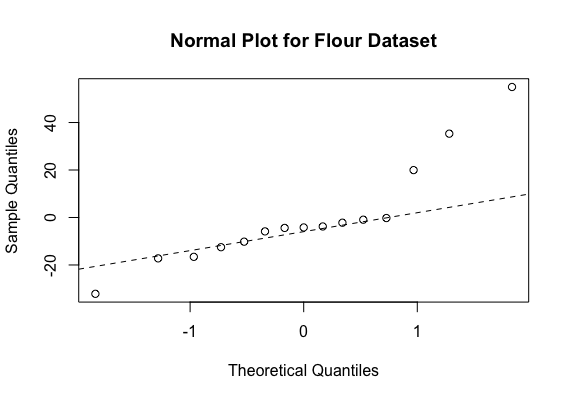
Regression Equation:

y = 0.021795x – 5.8643573

5.

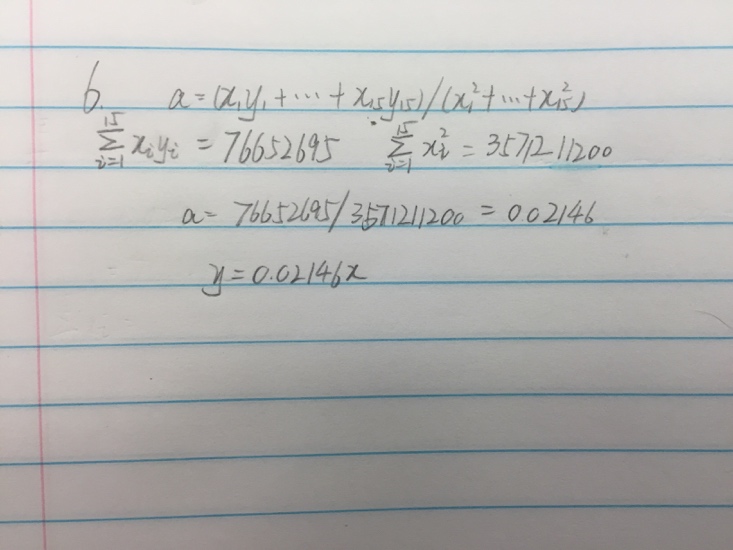


This residual plot is biased and heteroscedastic. And the pattern of residuals is random, so the relationship between NBags and weight is linear.



This normal plot is skewed to the right. It means the sample quantiles and theoretical quantiles are not linear.

6.



7.

Call:

lm(formula = NBags ~ Weight + 0, data = flour)

Residuals:

Min 1Q Median 3Q Max

-29.840 -13.118 -7.224 -2.360 54.429

Coefficients:

Estimate Std. Error t value Pr(>|t|)

Weight 0.0214641 0.0003673 58.43 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.95 on 14 degrees of freedom

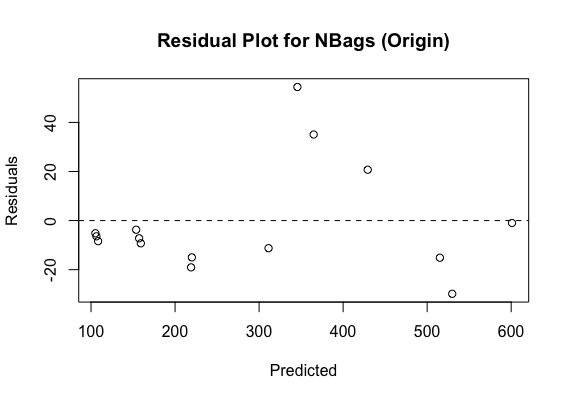
Multiple R-squared: 0.9959, Adjusted R-squared: 0.9956

F-statistic: 3414 on 1 and 14 DF, p-value: < 2.2e-16

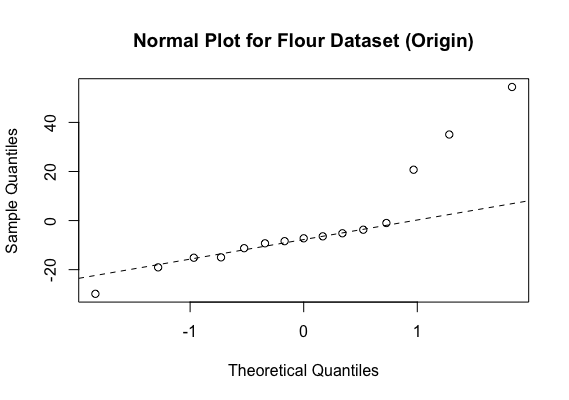
Regression Equation through origin:

y = 0.0214641x

8.



This residual plot is biased and heteroscedastic. And the pattern of residuals is random, so the relationship between NBags and weight is linear.



This normal plot is skewed to the right. It means the sample quantiles and theoretical quantiles are not linear.

Part B: (Data from www.truecar.com)

1.

Price Year Miles

1 11477 2000 76500

2 4933 2000 201950

3 11497 2001 47460

4 7996 2002 190580

5 13994 2003 128175

6 9795 2003 122780

7 11500 2004 131347

8 13425 2004 106430

9 15921 2005 80750

10 16777 2005 55385

11 16997 2006 88352

12 17900 2006 68013

13 15991 2007 69393

14 13900 2007 141890

15 22500 2007 61752

16 15995 2008 91147

17 20795 2008 57439

18 21995 2009 78090

19 19599 2010 80995

20 24900 2010 58203

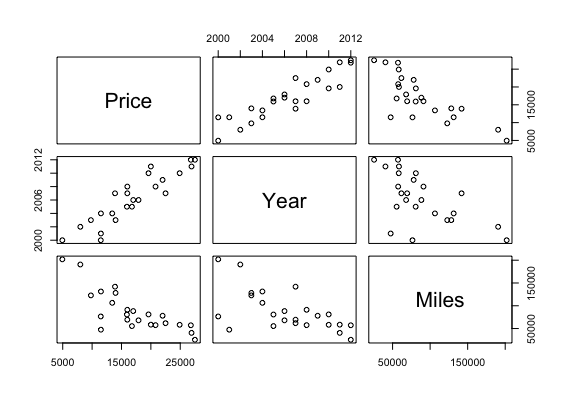
21 26921 2011 40486

22 19999 2011 58486

23 26804 2012 57210

24 27498 2012 25359

2.



3.

Price Year Miles

Price 1.0000000 0.9056820 -0.7862266

Year 0.9056820 1.0000000 -0.6080819

Miles -0.7862266 -0.6080819 1.0000000

Regression Equation:

y = 0.9056820x1 – 0.7862266x2

4.

Call:

lm(formula = Price ~ Year, data = UsedCars)

Residuals:

Min 1Q Median 3Q Max

-4208.2 -2391.2 849.8 1885.5 4391.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2990846.6 300191.6 -9.963 1.29e-09 \*\*\*

Year 1499.2 149.6 10.020 1.16e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2647 on 22 degrees of freedom

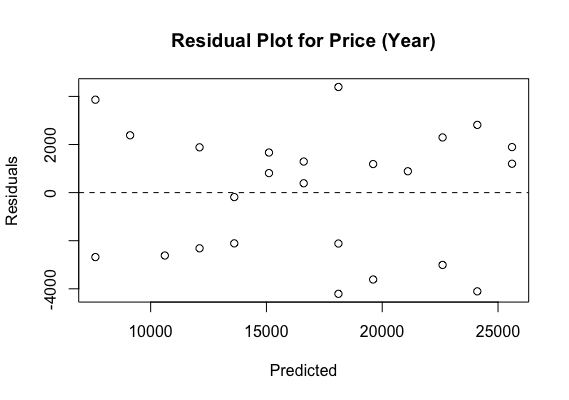
Multiple R-squared: 0.8203, Adjusted R-squared: 0.8121

F-statistic: 100.4 on 1 and 22 DF, p-value: 1.164e-09

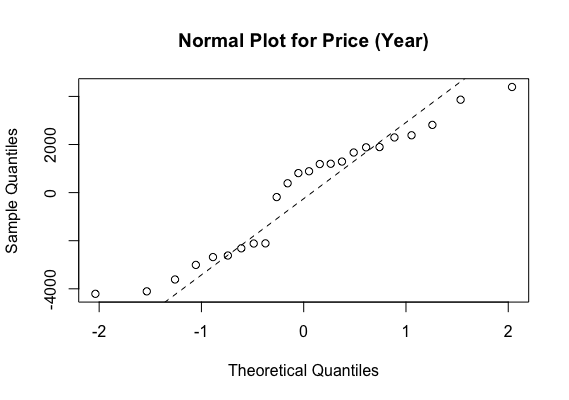
Regression Equation:

y = 1499.2x – 2990846.6

5.



This residual plot is unbiased and homoscedastic. And the pattern of residuals is random, so the relationship between year and price is linear.



This normal plot is thin tails. It means the sample quantiles and theoretical quantiles are not linear.

6.

Call:

lm(formula = Price ~ Miles, data = UsedCars)

Residuals:

Min 1Q Median 3Q Max

-9940.4 -1732.2 42.5 2607.3 6416.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.655e+04 1.776e+03 14.947 5.27e-13 \*\*\*

Miles -1.076e-01 1.804e-02 -5.968 5.25e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3858 on 22 degrees of freedom

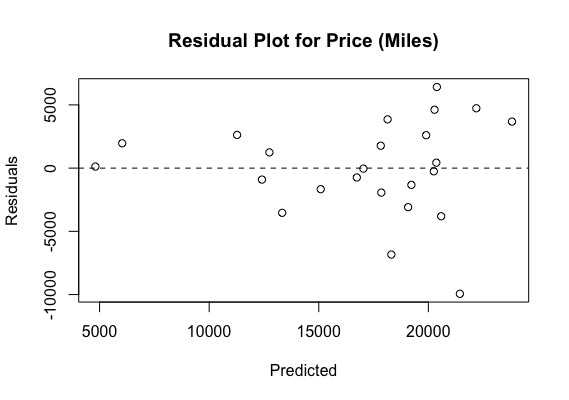
Multiple R-squared: 0.6182, Adjusted R-squared: 0.6008

F-statistic: 35.61 on 1 and 22 DF, p-value: 5.254e-06

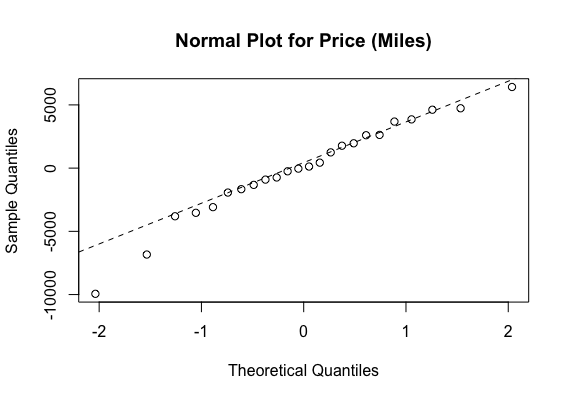
Regression Equation:

y = -1.076e-01x + 2.655e+04

7.



This residual plot is unbiased and homoscedastic.



This normal plot is skewed to the left. It means the sample quantiles and theoretical quantiles are not linear.

8.

Call:

lm(formula = Price ~ Year + Miles, data = UsedCars)

Residuals:

Min 1Q Median 3Q Max

-3858.1 -1321.9 199.7 1099.8 3302.4

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.232e+06 2.771e+05 -8.053 7.41e-08 \*\*\*

Year 1.123e+03 1.378e+02 8.149 6.12e-08 \*\*\*

Miles -5.115e-02 1.140e-02 -4.488 0.000202 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1936 on 21 degrees of freedom

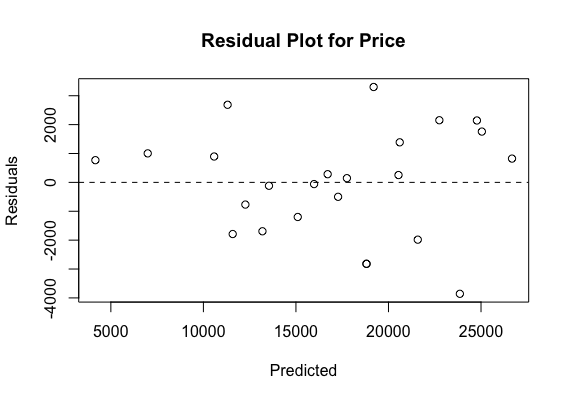
Multiple R-squared: 0.9083, Adjusted R-squared: 0.8995

F-statistic: 104 on 2 and 21 DF, p-value: 1.279e-11

Regression Equation:

y = 1.123e+03x1 – 5.115e-02x2 – 2.232e+06

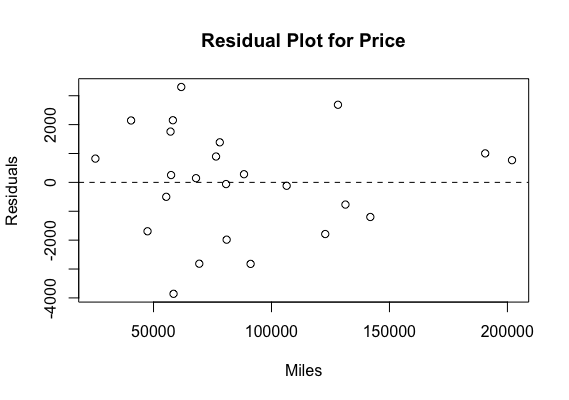
9.



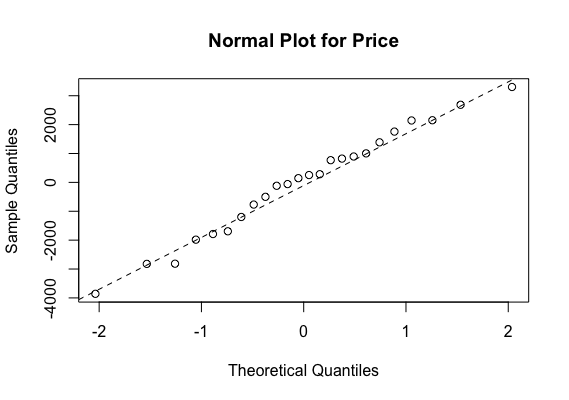
This residual plot is unbiased and homoscedastic. The pattern of residuals is random, so the relationship between predicted value and price is linear.



This residual plot is unbiased and homoscedastic. The pattern of residuals is random, so the relationship between year and price is linear.



This residual plot is unbiased and homoscedastic. And the pattern of residuals is random, so the relationship between miles and price is linear.



From this normal plot, all pattern is in a straight line, so the residuals are normally distributed.

10.

After comparing these regression models, in my view, the best regression model is the multiple linear regression, since in the multiple linear regression, there are two elements year and miles, which both influence the price. However, the simple linear regression only talks about one element each time, since price is not only influenced by one element, simple linear regression could not have enough power to provide a relatively accurate result compare with multiple linear regression. So as a conclusion, price = year miles is much better than the other two simple linear regressions.