Stat 448, homework 3
Shuhui Guo

Exercise 1

a)

			MPG (Highway)		
			Mean	Std	N
Cylinders	**Origin**	**Type**			
4	**Asia**	**Sedan**	33.35	4.27	49
		Sports	27.88	3.18	8
	USA	**Sedan**	32.69	3.31	29
6	**Asia**	**Sedan**	26.56	1.84	41
		Sports	26.33	1.51	6
	USA	**Sedan**	27.27	2.90	45
		Sports	27.00	2.83	2

According to the above table, the average highway fuel efficiency of cars with 4 cylinders is higher than that of cars with 6 cylinders. For each type of cars with 4 cylinders, the average highway fuel efficiency of cars from Asia is higher than that of cars from the USA. While for each type of cars with 6 cylinders, the average highway fuel efficiency of cars from Asia is lower than that of cars from the USA. Also, the average highway fuel efficiency of Sedan type is higher than that of Sports type. These cells do not have the same standard deviations but the differences in value are not large. And these cells do not have the same counts, thus it is an unbalanced dataset.

b)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1587.409567	529.136522	49.47	<.0001
Error	176	1882.651545	10.696884		
Corrected Total	179	3470.061111			

R-Square	Coeff Var	Root MSE	MPG_Highway Mean
0.457459	11.03900	3.270609	29.62778

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cylinders	1	1470.787732	1470.787732	137.50	<.0001
Origin	1	8.564346	8.564346	0.80	0.3721
Type	1	108.057489	108.057489	10.10	0.0018

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cylinders	1	1453.170429	1453.170429	135.85	<.0001
Origin	1	0.841224	0.841224	0.08	0.7795
Type	1	108.057489	108.057489	10.10	0.0018

According to the above three-way main effects ANOVA results, the p-value of this model is less than 0.0001, which indicates that the model is significant. The value of R-square is 0.457459, which means that 45.7459% of the variation in **mpg_highway** is explained by the model. According to Type I and Type III sum of squares, **Cylinders** and **Type** are significant because their p-values are less than 0.05. But **Origin** is not significant because its p-value is more than 0.05. Therefore, **Cylinders** and **Type** should be kept in a model for **mpg_highway**.

Results of the model with the predictors I decide to keep are as below:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1586.568342	793.284171	74.55	<.0001
Error	177	1883.492769	10.641202		
Corrected Total	179	3470.061111			

R-Square	Coeff Var	Root MSE	MPG_Highway Mean
0.457216	11.01023	3.262086	29.62778

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cylinders	1	1470.787732	1470.787732	138.22	<.0001
Type	1	115.780611	115.780611	10.88	0.0012

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cylinders	1	1481.993512	1481.993512	139.27	<.0001
Type	1	115.780611	115.780611	10.88	0.0012

According to the above results, the p-value of this model is less than 0.0001, which indicates that the model is significant. Based on Type I and Type III sum of squares, **Cylinders** and **Type** are both significant because their p-values are less than 0.05. The value of R-square is 0.457216, which means that 45.7216% of the variation in highway fuel efficiency is explained by the model.

c)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1670.425229	556.808410	54.45	<.0001
Error	176	1799.635883	10.225204		
Corrected Total	179	3470.061111			

R-Square	Coeff Var	Root MSE	MPG_Highway Mean
0.481382	10.79287	3.197687	29.62778

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cylinders	1	1470.787732	1470.787732	143.84	<.0001
Type	1	115.780611	115.780611	11.32	0.0009
Cylinders*Type	1	83.856886	83.856886	8.20	0.0047

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cylinders	1	207.5516175	207.5516175	20.30	<.0001
Type	1	116.6363540	116.6363540	11.41	0.0009
Cylinders*Type	1	83.8568863	83.8568863	8.20	0.0047

According to the above results, the p-value of this model is less than 0.0001, which indicates that the model is significant. The value of R-square is 0.481382, which means that 48.1382% of the variation in **mpg_highway** is explained by the model. According to Type I and Type III sum of squares, the two main effects and the interaction term between **Cylinders** and **Type** are all significant because their p-values are all less than 0.05.

The group differences are as below:

Cylinders	MPG_Highway LSMEAN	H0:LSMean1=LSMean2
		Pr > |t|
4	30.4887821	<.0001
6	26.7151163	

Least Squares Means for Effect Cylinders				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	**2**	3.773666	2.120634	5.426697

Type	MPG_Highway LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
Sedan	30.0163983	0.0009
Sports	27.1875000	

Least Squares Means for Effect Type				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	2.828898	1.175867	4.481930

Cylinders	Type	MPG_Highway LSMEAN	LSMEAN Number
4	Sedan	33.1025641	1
4	Sports	27.8750000	2
6	Sedan	26.9302326	3
6	Sports	26.5000000	4

Least Squares Means for Effect Cylinders*Type				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	5.227564	2.148489	8.306639
1	3	6.172332	4.875485	7.469178
1	4	6.602564	3.523489	9.681639
2	3	0.944767	-2.120956	4.010491
2	4	1.375000	-2.771993	5.521993
3	4	0.430233	-2.635491	3.495956

According to above results, the average highway fuel efficiency difference between 4 cylinders and 6 cylinders is 3.773666. And the corresponding 95% confidence interval is (2.120634, 5.426697), which does not include 0. Also, the p-value is less than 0.0001, which rejects the null hypothesis that there is no difference between these two groups. Therefore, it is believed that the average highway fuel efficiency of cars with 4 cylinders is significantly higher than that of cars with 6 cylinders.

The average highway fuel efficiency difference between Sedan type and Sports type is 2.828898. And the corresponding 95% confidence interval is (1.175867, 4.481930), which does not include 0. Also, the p-value is less than 0.05, which rejects the null hypothesis that there is no difference between these two groups. Therefore, we can say that the average highway fuel efficiency of Sedan type is significantly higher than that of Sports type.

For interaction term, the 95% confidence intervals for the difference between Sedan type with 4 cylinders and Sports type with 4 cylinders, the difference between Sedan type with 4 cylinders and Sedan type with 6 cylinders, the difference between Sedan type with 4 cylinders and Sports type with 6 cylinders do not include 0. And the difference estimates are all positive. Therefore, we can say that the average highway fuel efficiency of Sedan type with 4 cylinders is significantly higher than that of Sports type with 4 cylinders. The average highway fuel efficiency of Sedan type with 4 cylinders is significantly higher than that of Sedan type with 6 cylinders. The average highway fuel efficiency of Sedan type with 4 cylinders is significantly higher than that of Sports type with 6 cylinders.

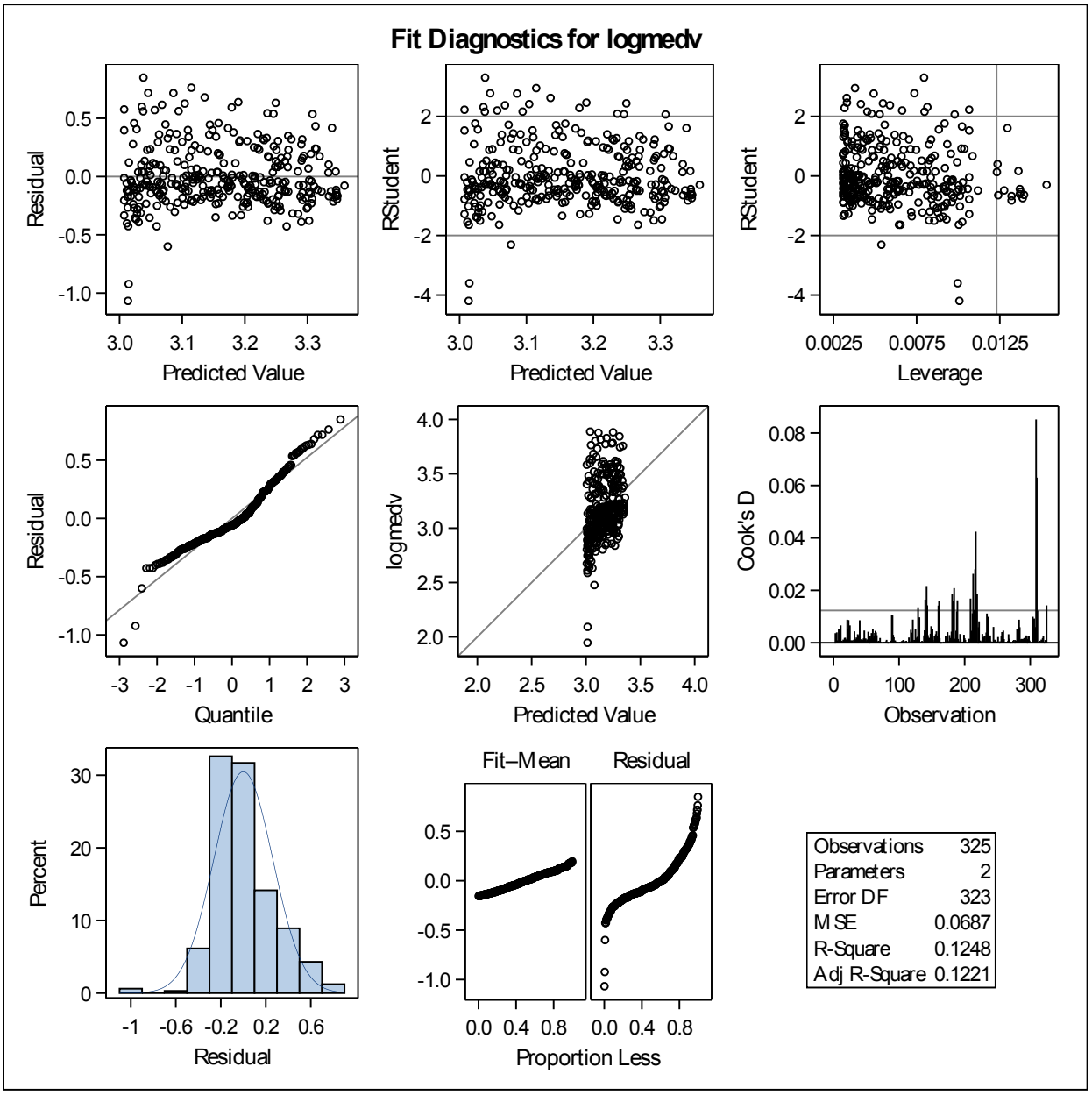
Exercise 2

a)

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	3.16665	3.16665	46.08	<.0001
Error	323	22.19838	0.06873		
Corrected Total	324	25.36503			

Root MSE	0.26216	**R-Square**	0.1248
Dependent Mean	3.16225	**Adj R-Sq**	0.1221
Coeff Var	8.29017		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	3.36960	0.03383	99.60	<.0001
age	1	-0.00362	0.00053373	-6.79	<.0001



According to the simple linear regression results, the p-value of this model is less than 0.0001, which indicates that this model is significant. Also, the p-values of parameters are less than 0.0001, which means that the intercept and age term are both significant. The R square is 0.1248, which means only 12.48% of the variation in **logmedv** is explained by this model.

The diagnostics of this simple linear regression could be analyzed in the following aspects:

i. The plots of residual and studentized residual versus predicted value are not evenly distributed vertically. The negative points are much more than positive points. And there are outliers in the two plots.

ii. The plot of studentized residual versus leverage seems to indicate that there are some outlying data points. The plot of Cook's distance versus observation number also reveals that there are some data points

above the cutoff line. To address this issue, the data point which has Cook's distance greater than 4 times the cutoff line will be removed while constructing the final model.

iii. The point pattern in the normal quantile plot of residuals does not appear to be a straight line. Also, the residual histogram seems not fit the curve of normal distribution. But the residuals appear to follow the short-tailed distribution, thus this issue does not require remedies necessarily.

iv. The points in the plot of the dependent variable versus the predicted value do not follow the 45-degree line, which indicates that this model is not a perfect fit.

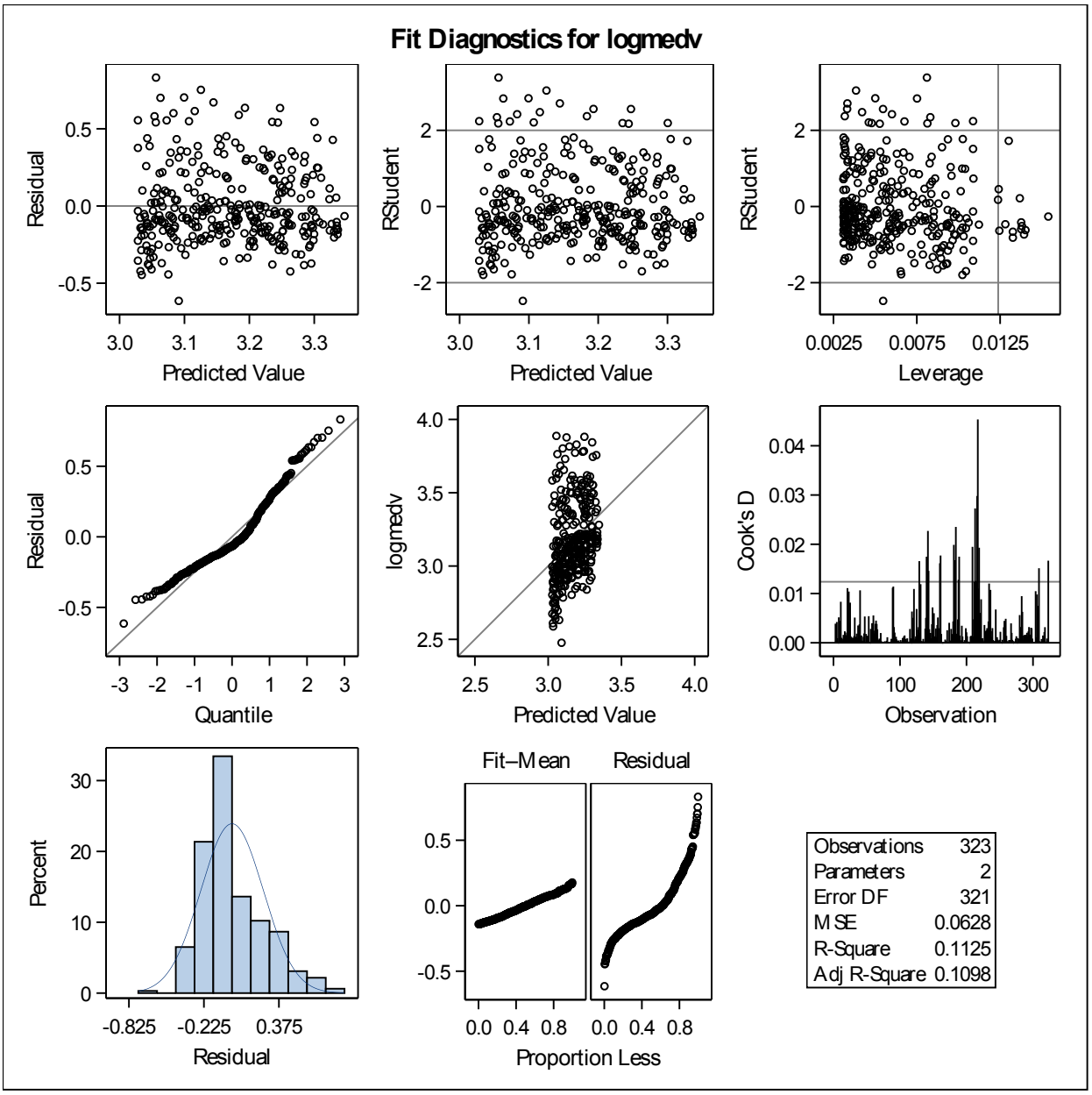
v. In the Residual-Fit Spread Plot, the right side (residuals) is taller than the left side (fit), which means that the spread of the residuals is greater than the spread of the fit. There is still a lot of unexplained variation.

After removing the high influential points and re-fitting the model, the results are shown as below:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	2.55678	2.55678	40.70	<.0001
Error	321	20.16686	0.06283		
Corrected Total	322	22.72364			

Root MSE	0.25065	**R-Square**	0.1125
Dependent Mean	3.16933	**Adj R-Sq**	0.1098
Coeff Var	7.90859		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	3.35613	0.03243	103.48	<.0001
age	1	-0.00328	0.00051391	-6.38	<.0001



b)

According to the above results, the p-value of this model is less than 0.0001, which indicates that this model is significant. Also, the p-values of parameters are less than 0.0001, which means that the intercept and age term are both significant. The estimated coefficient of **age** is -0.00328, and exp(-0.00328)=0.99673. Thus an increase of home age is associated with a decrease of median home value. The average median home value is predicted to have a 0.99673 percent change when home age increases by 1 unit. The R square is 0.1125, which means only 11.25% of the variation in **logmedv** is explained by this model.

In the final model, the data points with high influence are removed. But there are some remaining issues in the diagnostics:

i. The points in the plot of the dependent variable versus the predicted value do not follow the 45-degree line, which indicates that this model is not a perfect fit.

ii. In the Residual-Fit Spread Plot, the right side (residuals) is taller than the left side (fit), which means that the spread of the residuals is greater than the spread of the fit. There is still a lot of unexplained variation.

These issues could be remedied by adding predictors and fitting an adequate model to estimate median home value.

Because the variation explained by this model is too small and there are some remaining issues in the diagnostics, the model based on age alone is not much useful for estimating median home value.

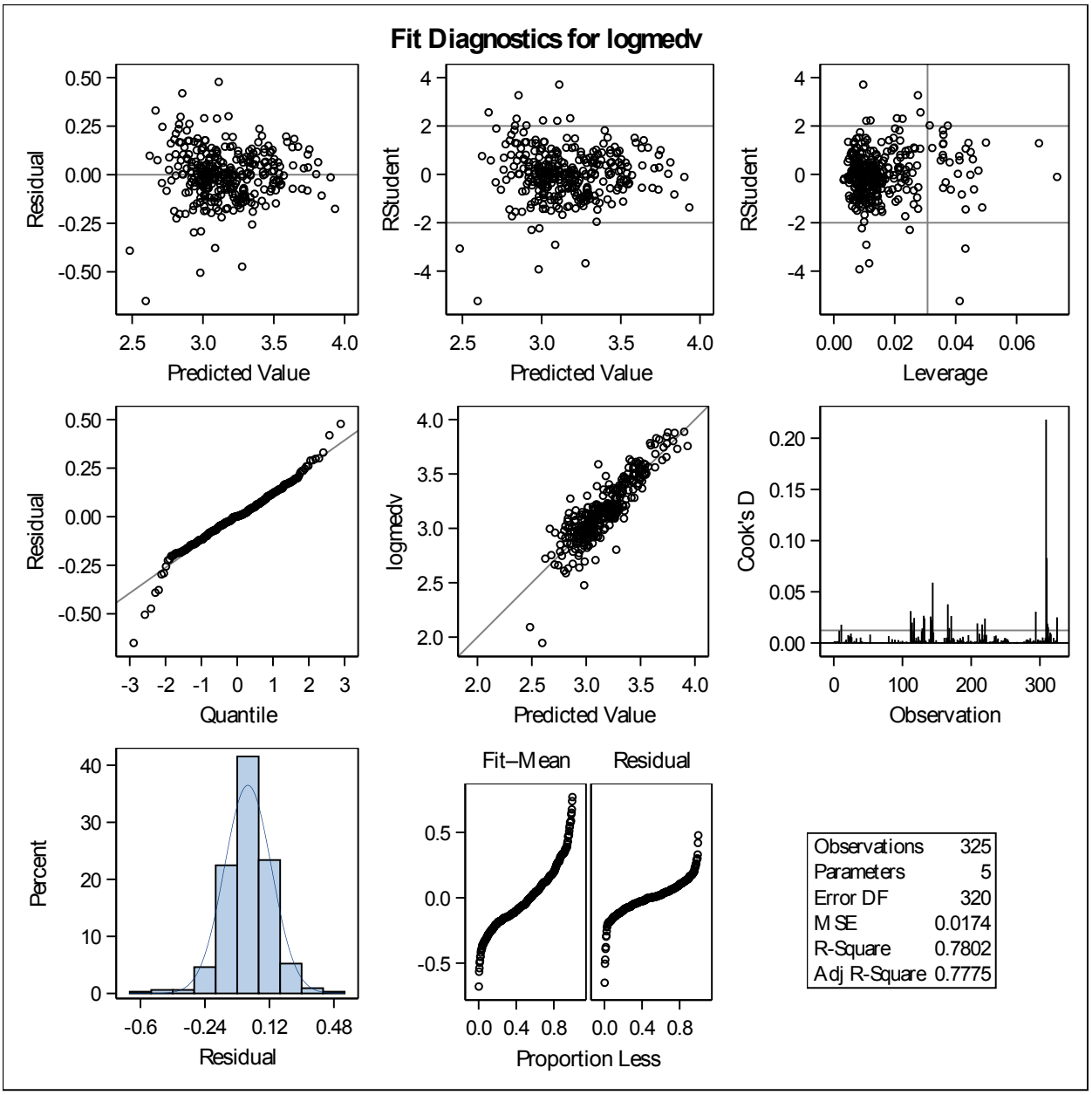
Exercise 3

a)

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	19.79088	4.94772	284.04	<.0001
Error	320	5.57415	0.01742		
Corrected Total	324	25.36503			

Root MSE	0.13198	**R-Square**	0.7802
Dependent Mean	3.16225	**Adj R-Sq**	0.7775
Coeff Var	4.17367		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	0.87577	0.11130	7.87	<.0001	0
age	1	-0.00238	0.00038830	-6.13	<.0001	2.08818
indus	1	-0.00896	0.00171	-5.24	<.0001	1.78209
nox	1	0.46342	0.17909	2.59	0.0101	2.58044
rm	1	0.35470	0.01349	26.29	<.0001	1.19781



According to the linear regression results, the p-value of this model is less than 0.0001, which indicates that this model is significant. Also, the p-values of parameters are less than 0.05, which means that the intercept and coefficients are all significant.

The diagnostics of this linear regression could be analyzed in the following aspects:

i. There are outliers in the plots of residual and studentized residual versus predicted value.

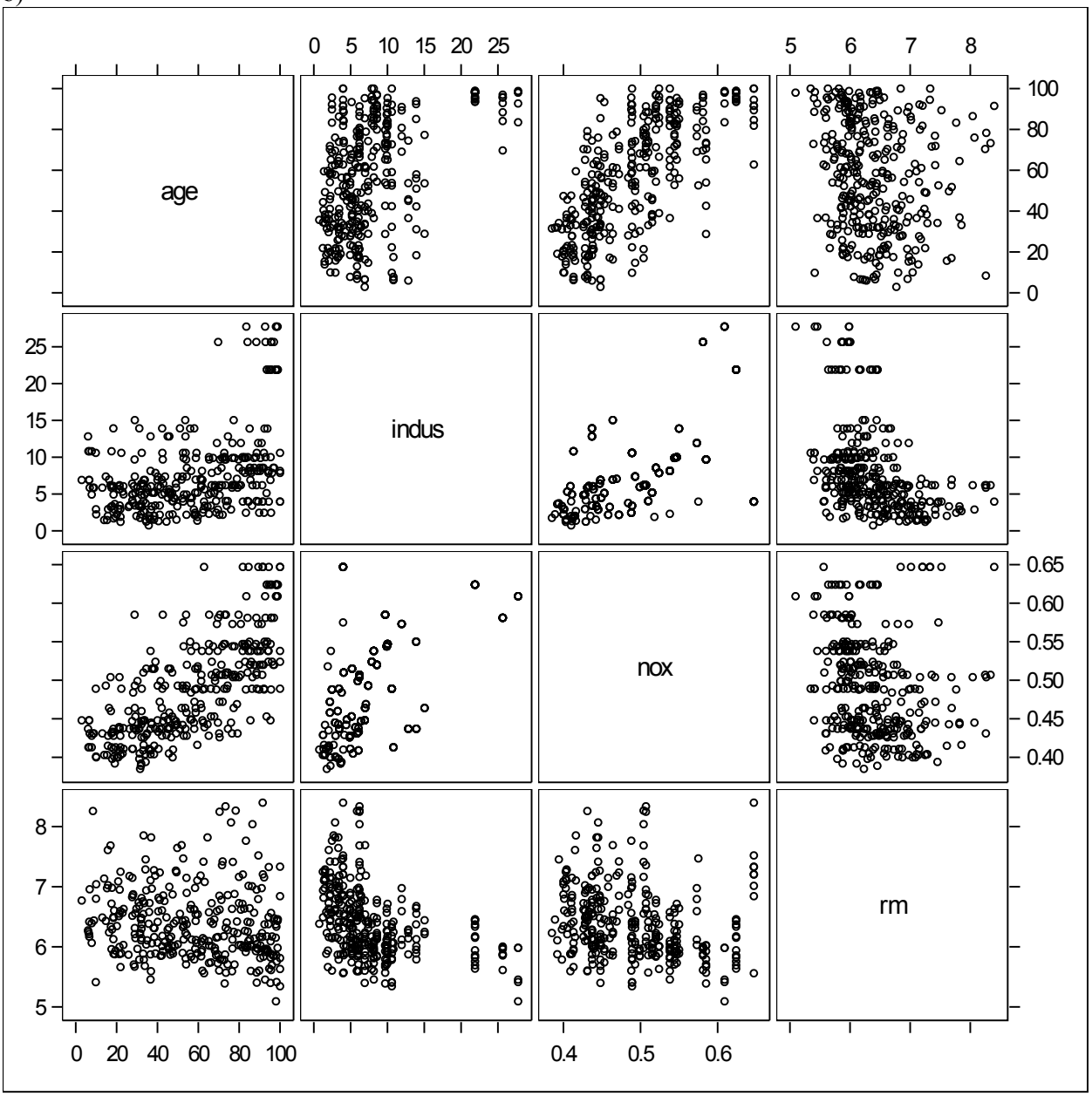
ii. The plot of studentized residual versus leverage seems to indicate that there are some outlying data points. The plot of Cook's distance versus observation number also reveals that there are some data points above the cutoff line.

iii. The point pattern in the normal quantile plot of residuals is roughly a straight line except some outliers. But the residuals appear to follow the short-tailed distribution, thus this issue does not require remedies necessarily.

iv. Most of the points in the plot of the dependent variable versus the predicted value are in the 45-degree line. There are some outliers.

v. In the Residual-Fit Spread Plot, the left side (fit) is taller than the right side (residuals), which means that this model explains a lot of variation in **logmedv**.

b)



The R square is 0.7802 (adjusted R square is 0.7775), which means 78.02% of the variation in **logmedv** is explained by this model.

The p-values of parameters are all less than 0.05, which means that the intercept and coefficients are all significant. The estimated coefficient of **age** is -0.00238. When other variables are fixed, an increase of home age is associated with a decrease of **logmedv**. On average, **logmedv** is predicted to have a decrease of 0.00238 units when home age increases by 1 unit. The estimated coefficient of **indus** is -0.00896. When other variables are fixed, an increase of the proportion of non-retail business acres is associated with a decrease of **logmedv**. On average, **logmedv** is predicted to have a decrease of 0.00896 units when the proportion of non-retail business acres increases by 1 unit. The estimated coefficient of **nox** is 0.46342. When other variables are fixed, an increase of nitric oxides concentration is associated with an increase of **logmedv**. On average, **logmedv** is predicted to have an increase of 0.46342 units when nitric oxides concentration increases by 1 unit. The estimated coefficient of **rm** is 0.35470. When other variables are fixed, an increase of average number of rooms per house is associated with an increase of **logmedv**. On average, **logmedv** is predicted to have an increase of 0.35470 units when average number of rooms per house increases by 1 unit.

According to the scatterplot matrix, there are no strongly correlated predictors. In addition, the VIFs of predictors are all less than 10. Therefore, there is no multicollinearity issue among the predictors.

This model explains a lot of variations in the response variable. And there is no multicollinearity issue among predictors. But this model has some diagnostic issues needed to be fixed. Therefore, this is not a good model to predict median home value.

Exercise 4

a)

Backward Elimination: Step 0

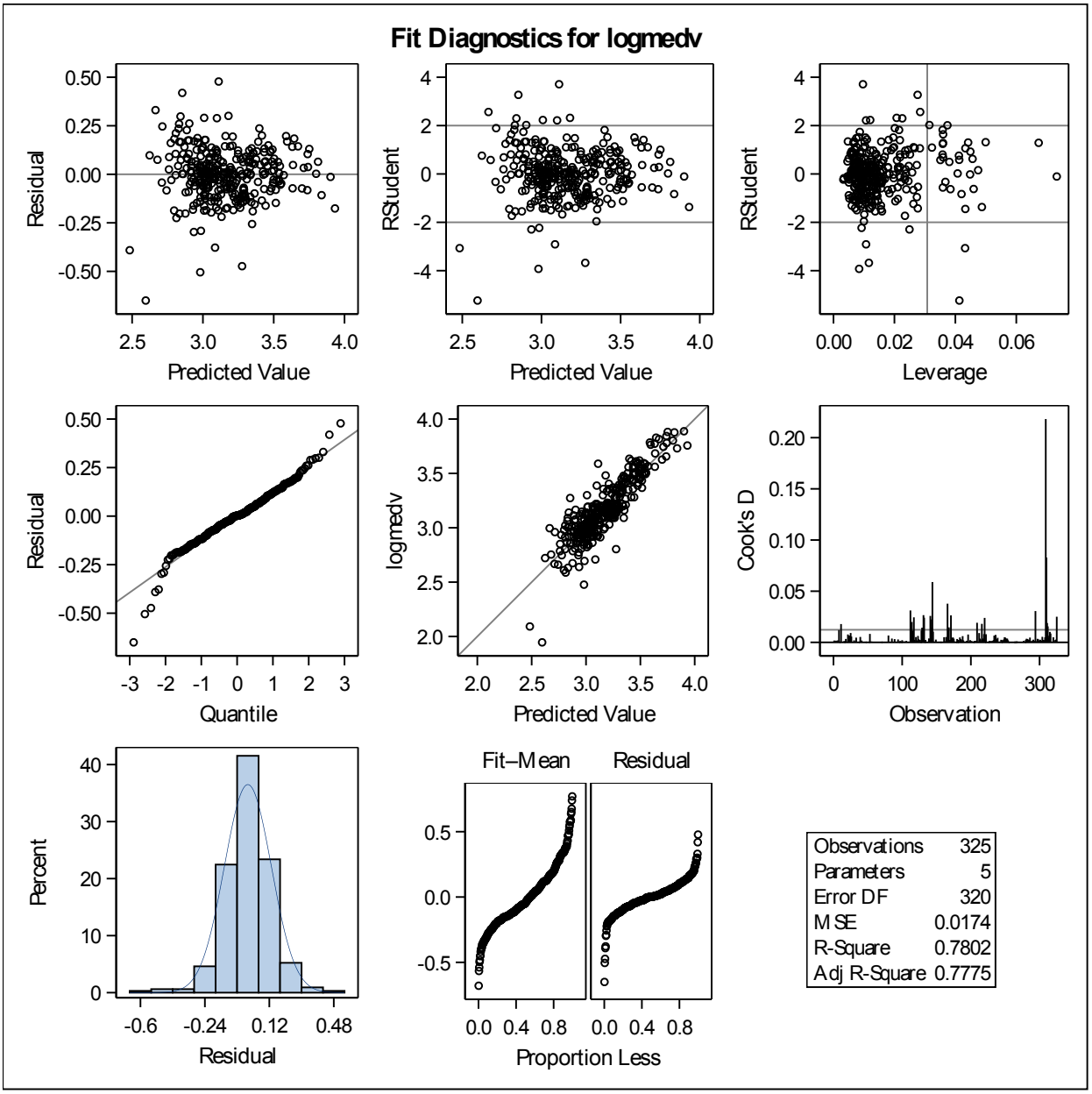
All Variables Entered: R-Square = 0.7802 and C(p) = 5.0000

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	19.79088	4.94772	284.04	<.0001
Error	320	5.57415	0.01742		
Corrected Total	324	25.36503			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	0.87577	0.11130	1.07856	61.92	<.0001
indus	-0.00896	0.00171	0.47909	27.50	<.0001
nox	0.46342	0.17909	0.11663	6.70	0.0101
rm	0.35470	0.01349	12.03927	691.15	<.0001
age	-0.00238	0.00038830	0.65362	37.52	<.0001

Bounds on condition number: 2.5804, 30.594

All variables left in the model are significant at the 0.0500 level.



After model selection, all the variables are left in the best linear regression model since they are all significant at the 0.05 level.

The diagnostics of this simple linear regression could be analyzed in the following aspects:

i. There are outliers in the plots of residual and studentized residual versus predicted value.

ii. The plot of studentized residual versus leverage seems to indicate that there are some outlying data points. The plot of Cook's distance versus observation number also reveals that there are some data points above the cutoff line. To address this issue, the data point which has Cook's distance greater than 4 times the cutoff line will be removed while constructing the final model.

iii. The point pattern in the normal quantile plot of residuals is roughly a straight line except some outliers. But the residuals appear to follow the short-tailed distribution, thus this issue does not require remedies necessarily.

iv. The points in the plot of the dependent variable versus the predicted value are in the 45-degree line except some outliers.

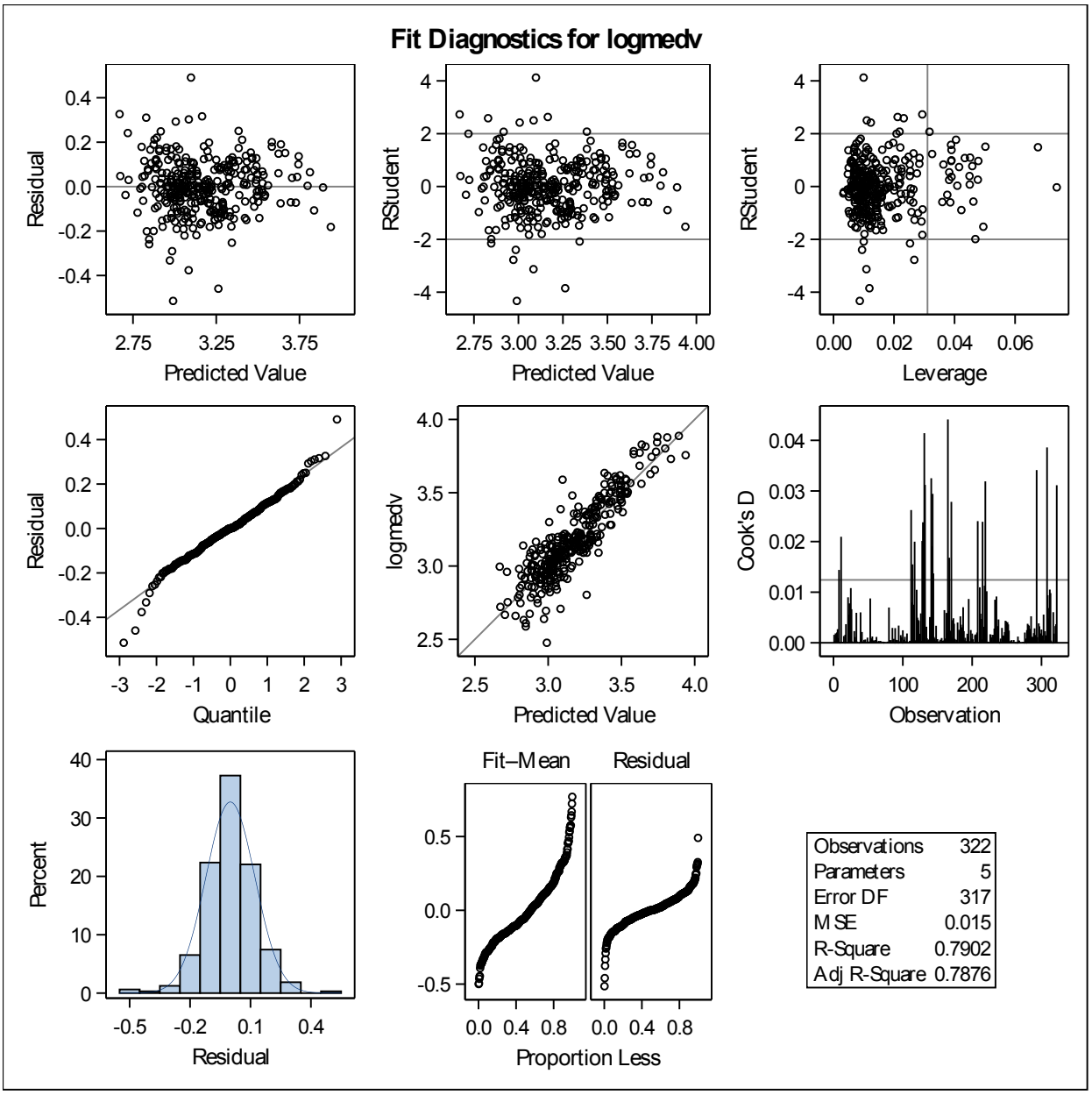
v. In the Residual-Fit Spread Plot, the left side (fit) is taller than the right side (residuals), which means that this model explains a lot of variation in **logmedv**.

Re-fit the model and the results are shown as below:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	17.94849	4.48712	298.56	<.0001
Error	317	4.76429	0.01503		
Corrected Total	321	22.71278			

Root MSE	0.12259	**R-Square**	0.7902
Dependent Mean	3.16901	**Adj R-Sq**	0.7876
Coeff Var	3.86853		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.85758	0.10397	8.25	<.0001
age	1	-0.00247	0.00036252	-6.82	<.0001
indus	1	-0.00624	0.00163	-3.83	0.0002
nox	1	0.45177	0.16650	2.71	0.0070
rm	1	0.35636	0.01262	28.25	<.0001



b)

In the final model, the data points with high influence are removed. There is no remaining issue in the diagnostics.

According to the above results, the p-value of this model is less than 0.0001, which indicates that this model is significant. Also, the p-values of parameters are less than 0.05, which means that the intercept and coefficients are all significant. The R square is 0.7902 (adjusted R square is 0.7876), which means 79.02% of the variation in **logmedv** is explained by this model.

The estimated coefficient of **age** is -0.00247, and exp(-0.00247)=0.99753. When other variables are fixed, an increase of home age is associated with a decrease of median home value. The average median home value is predicted to have a 0.99753 percent change when home age increases by 1 unit. The estimated

coefficient of **indus** is -0.00624, and exp(-0.00624)=0.99378. When other variables are fixed, an increase of the proportion of non-retail business acres is associated with a decrease of median home value. The average median home value is predicted to have a 0.99378 percent change when the proportion of non-retail business acres increases by 1 unit. The estimated coefficient of **nox** is 0.45177, and exp(0.45177)=1.57109. When other variables are fixed, an increase of nitric oxides concentration is associated with an increase of median home value. The average median home value is predicted to have a 1.57109 times change when nitric oxides concentration increases by 1 unit. The estimated coefficient of **rm** is 0.35636, and exp(0.35636)=1.42812. When other variables are fixed, an increase of average number of rooms per house is associated with an increase of median home value. The average median home value is predicted to have a 1.42812 times change when average number of rooms per house increases by 1 unit.