Exercise1

(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

From the tabulation, apparently we see that the data is not balanced, so we will need to use proc glm to perform ANOVA test. In terms of mean of MPG_highway, it looks like 4 cylinder cars may be more fuel efficient than 6 cylinders. Sedans might be more efficient than sports cars at least for 4 cylinders. There does not appear to be a big difference across origin.

(b) To find the best main effects ANOVA model for mpg_highway, we first fit 3-way ANOVA model with cylinders, origin, and type. Type1 and Type3 results are as follows.

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

Both Type1 and Type3 analyses show that Origin is not significant variable in ANOVA model. We remove Origin and re-fit the 2-way main effects ANOVA model with Cylinders and Type. Now, as we can see from the output below, Cylinders and Type are significant with small p-values in both Type1 and Type3 analyses. In other words, we should keep two main effects in our final model.

This 2-way main effects ANOVA model is highly significant with very small p-values, and both Cylinders and Type are highly significant. The model describes 45.7% of the variation in highway fuel efficiency.

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

(c) The Cylinders*Type interaction is also significant when added to the model, so the final model has Cylinders, Type, and their interaction. This 2-way ANOVA model with interaction is highly significant with small p-values and the model describes 48.14% of the variation in fuel efficiency.

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

The following differences of least squares means for main effects tell us that 4 cylinder cars are significantly more efficient than 6 cylinder cars, with 4 cylinders expected to get 3.77 mpg more than 6 cylinders and a confidence interval of (2.12, 5.43) for the difference. Sedans are significantly more efficient and expected to get 2.83 mpg more than sports cars with a confidence interval of (1.18, 4.48).

For the interaction term, only the vehicle, sedan with 4 cylinders, have significant differences with other type and 4cylinder combinations. It is significantly more fuel efficient than the other type and cylinder combinations with all confidence intervals being positive and not including zero.

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

		H0:LSMean1=LSMean2	
Cylinders	**MPG_Highway LSMEAN**	**t Value**	**Pr > |t|**
4	30.4887821	4.51	<.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

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

		H0:LSMean1=LSMean2	
Type	**MPG_Highway LSMEAN**	**t Value**	**Pr > |t|**
Sedan	30.0163983	3.38	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

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

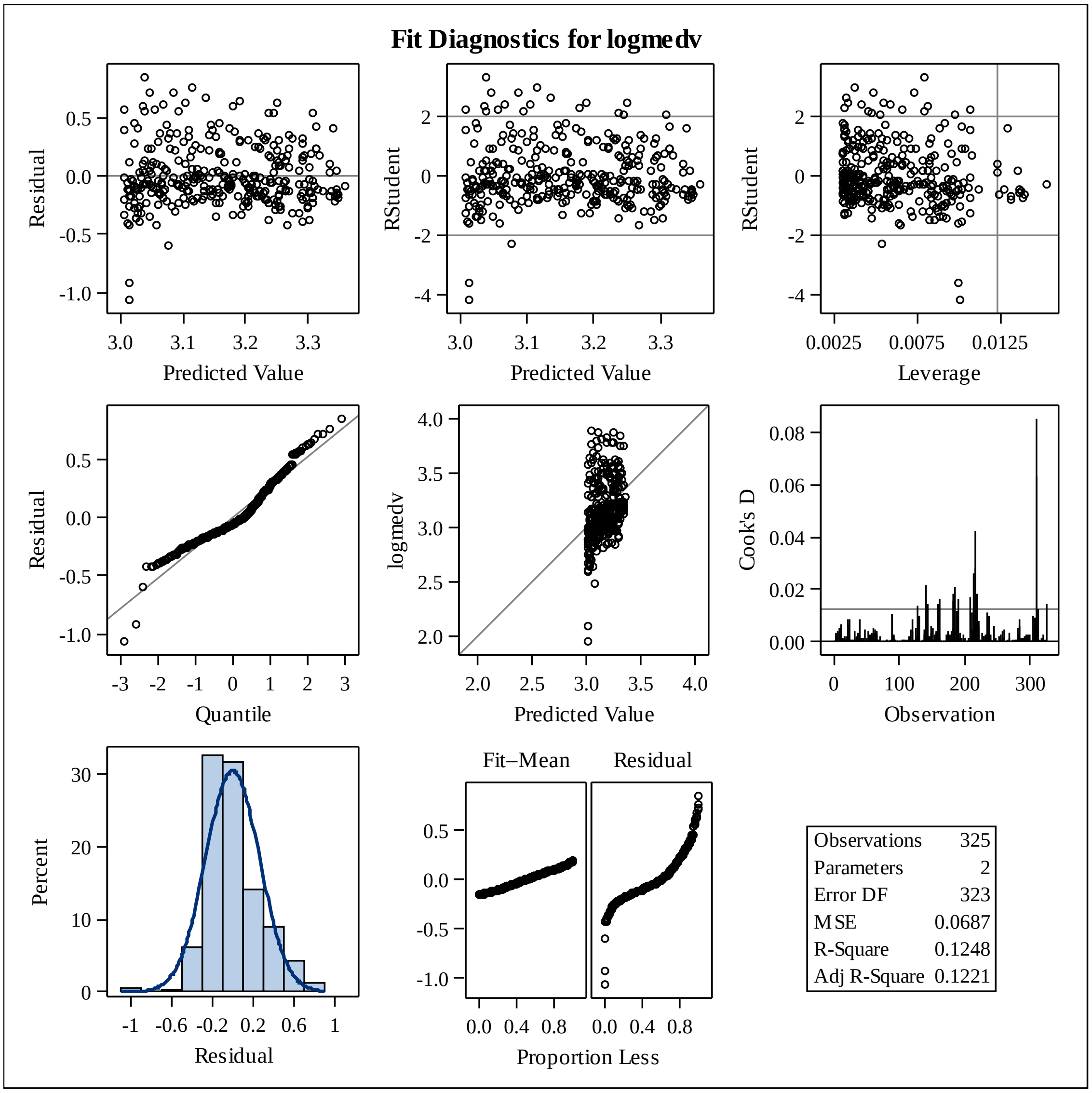
Cylinders	Type	MPG_Highway LSMEAN	LSMEAN Number
4	Sedan	33.1025641	1
4	Sports	27.8750000	2
6	Sedan	26.9302326	3
6	Sports	26.5000000	4

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

Exercise2

(a) We fit a simple linear regression model of logmev as a function of age for suburbs with less than 1 crime per capita. The following are the diagnostic plots. We should expect some Studentized residuals with absolute value greater than 2 given the number of observations. There are a couple close to -4 which might be problem. If those points have undue influence, we will remove them when we look at Cook's distances. Besides potential influential points, majority looks fine in residual plot. Also no pattern is detected in this plot.

For QQ plot, some points are not lying in the straight line, but histogram from residuals looks pretty symmetric and bell-shaped, thus normality assumption for error term looks valid. From the plot of Cook's D, some data show pretty large values compared to others, so we will delete points in the model that have Cook's distance greater than around 0.015*4=0.06. Technically, we should remove points one at a time and re-fit to re-check influence. In this particular case, the number of high influence points is small and the data is reasonably large, so we would be OK removing a couple of points at once. Since we do not detect serious problem with diagnostic plots, the same model is re-fitted after deleting some potentially unduly influential points.



Observations	325
Parameters	2
Error DF	323
MSE	0.0687
R-Square	0.1248
Adj R-Square	0.1221

(b) The simple regression model is significant with p-value less than .0001 and it can explain 11.25% of variation in log of median home value. The coefficient is estimated as -0.00328, which means that the change of expected median of house value is multiplied by exp(-0.00328)=0.9967 with one unit increase in age. It indicates a slight multiplicative decrease for one year increase in house age.

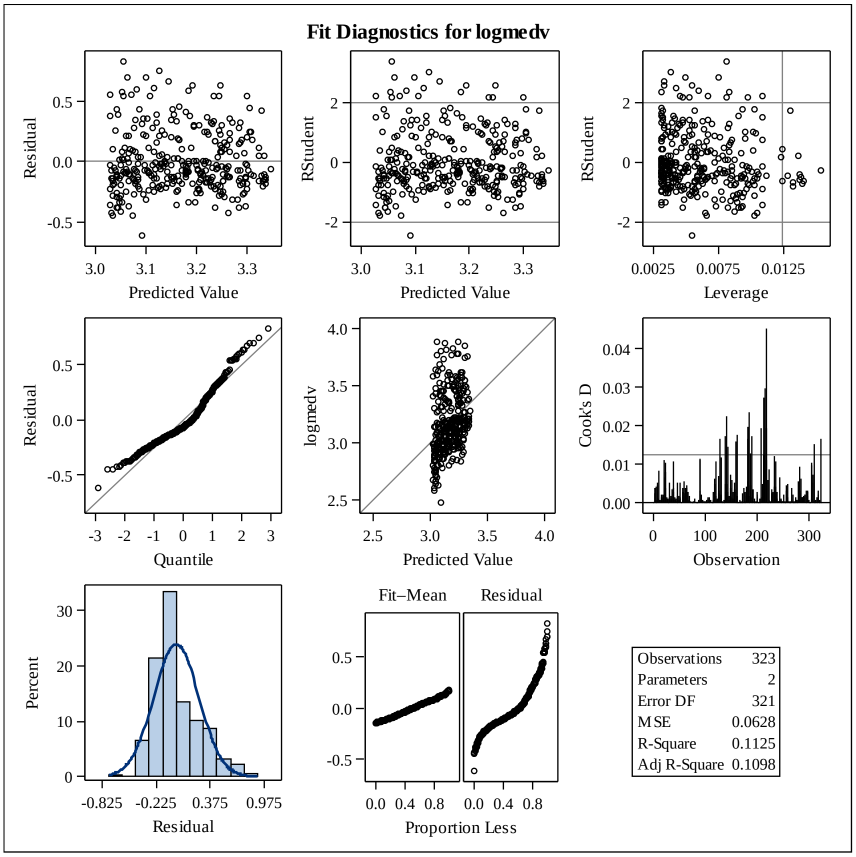
About the diagnostic plots, since we delete some potential influential observations, now residual plot and Cook's distance plot look better than the results from part (a). But we can still find some points with larger Cook's distance. There are within the 4*4/n limit prescribed in the assignment, so we will not remove further. Similar to the output in (a), normal QQ plot again shows some points not lying in the straight line, but it seems not serious.

The model with age alone can't explain much variation of the log median home value based on the small R square value. It would be better to include other useful predictors to increase the prediction power.

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			

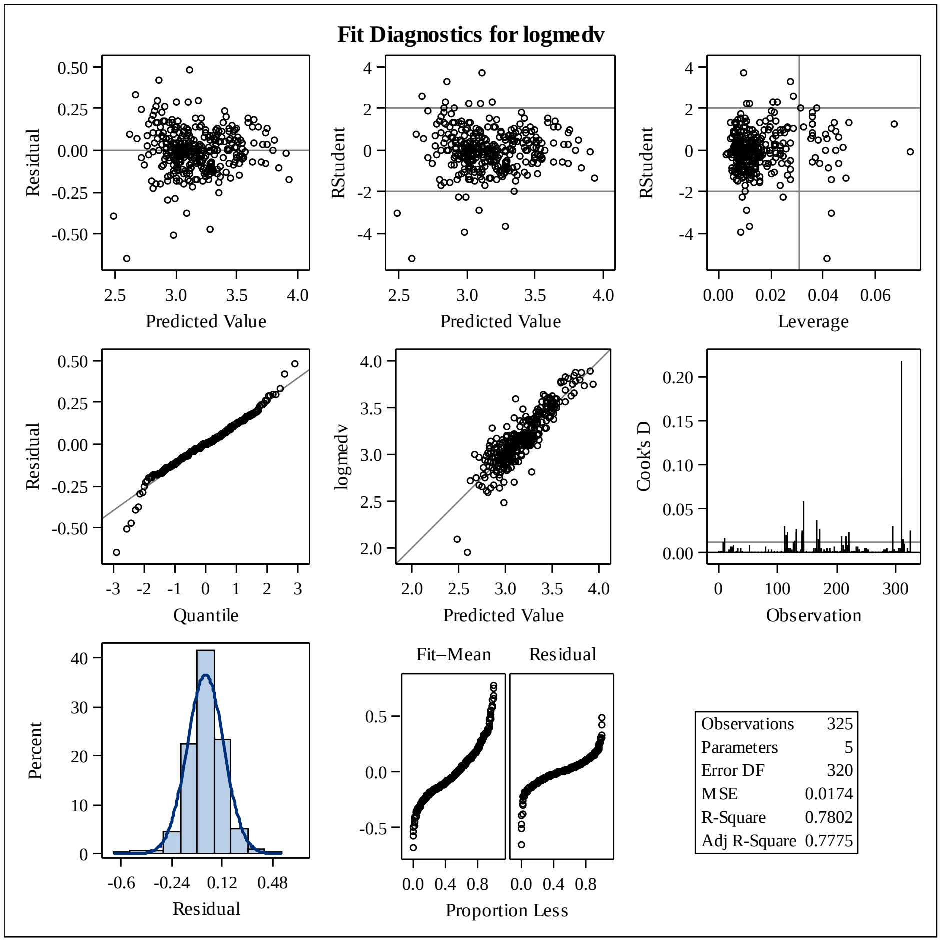
Root MSE	0.25065	**R-Square**	0.1125
Dependent Mean	3.16933	**Adj R-Sq**	0.1098
Coeff Var	7.90859		

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



Exercise3

(a) The diagnostics plot from a linear regression model for logmev with age, indus, nox, and rm shows that there exists some potential influential points with large Cook's distance. They may lead to residuals plot with large absolute values of residuals and non-straight line in QQ-plot. The histogram looks like a symmetric bell shape, but it shows a bit long left tail. We need to refit the model after removing some suspicious outliers.



Observations	325
Parameters	5
Error DF	320
MSE	0.0174
R-Square	0.7802
Adj R-Square	0.7775

(b) The variation explained by this regression model is 78.02%. It is a huge improvement compared to R^2 from the simple regression model in Exercise 2 (b). The R^2 always increases as the number of predictors increases.

We can also see that all four predictors are significant with p-values close to zero. For the interpretation on the relationship between logmedv and four significant predictors, coefficients for age and indus are negative and estimated as -0.00238, -0.00896, respectively. It means that the expected log median of house value decreases by 0.00238 and 0.00896 with one unit increase in age and indus, respectively. The coefficients for nox and rm are positive and estimated as 0.4632 and 0.3547, which means that the expected log median of house value increases by 0.4632 and 0.3547 with one unit increase in nox and rm, respectively. This interpretation is for logmev and the relationship between median home value and predictors will be described in Exercise4 (b).

All four predictors have VIF less than 10, so there is no multicollinearity issue. Although this model explains a lot of variation in log median home value, it is not a good model due to potential outliers.

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

Exercise 4

(a) The summary of stepwise selection is shown in the table below. It suggests that all of the four variables are significant and should be kept in the model.

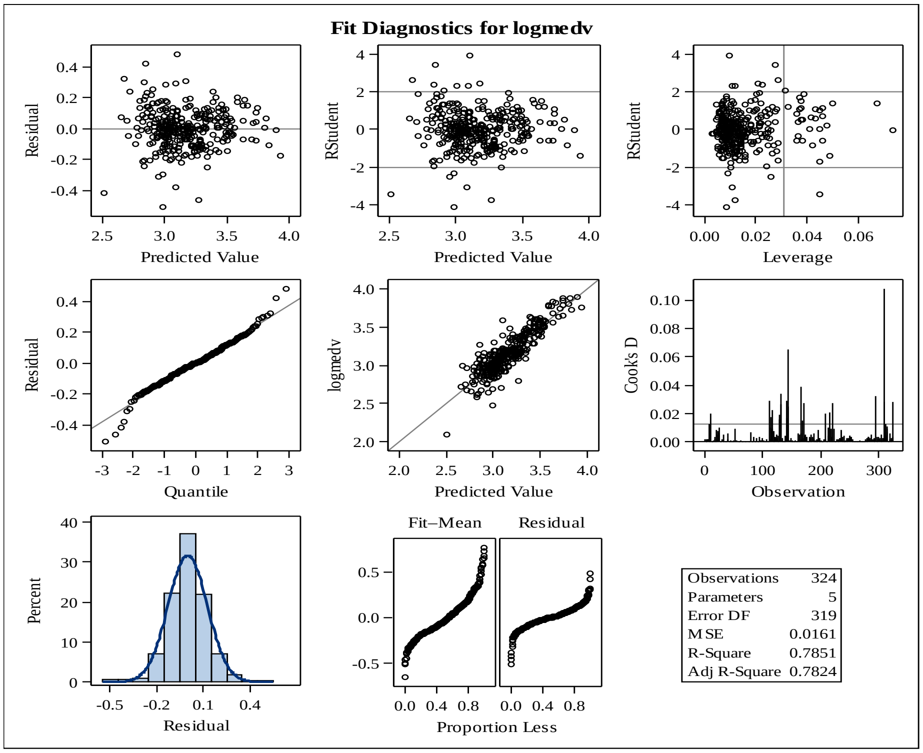
Model: MODEL1

Dependent Variable: logmedv

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	rm		1	0.7108	0.7108	100.156	793.78	<.0001
2	age		2	0.0505	0.7612	28.6575	68.07	<.0001
3	indus		3	0.0144	0.7756	9.6956	20.60	<.0001
4	nox		4	0.0046	0.7802	5.0000	6.70	0.0101

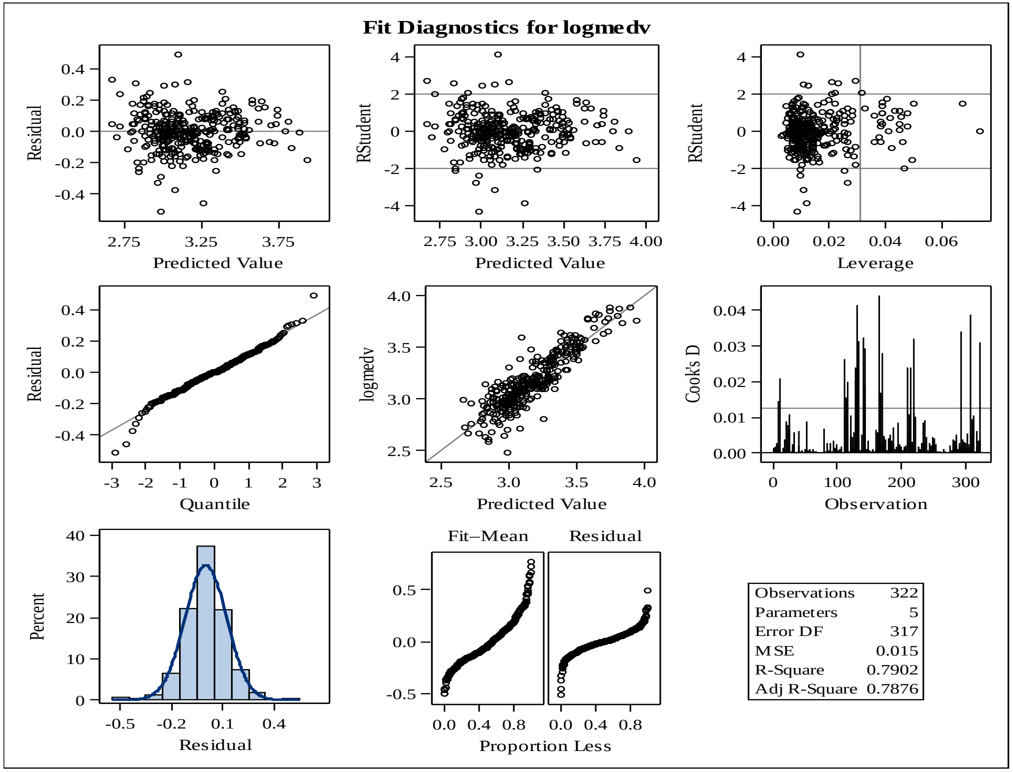
Indeed final regression model is just same as the model we run in Exercise 3. As we could see, there is an observation with standardized residual below -4, which may need a careful look. From the plot of Cook's D, some data show pretty large value compared to others, so we will delete points in the model that have Cook's distance greater than 0.1 first.

After deleting the observations with Cook's distance greater than 0.1, we refit the model and get the following diagnostic plots. Now residual plots look better but some data still show larger Cook's distance than others. So we will delete points in the model that have Cook's distance greater than around 0.015*4=0.06. Then the same model is re-fitted after deleting those potentially unduly influential points.



Observations	324
Parameters	5
Error DF	319
MSE	0.0161
R-Square	0.7851
Adj R-Square	0.7824

After deleting potential outliers, we re-fit the model again and check the diagnostics plot below. Now it there is no more observation showing Cook's distance larger than 4 times of cut-off line. Thus, total 322 observations are the final subset of original data (original size = 325) fitted to find a relationship between median home value and four predictors.



Observations	322
Parameters	5
Error DF	317
MSE	0.015
R-Square	0.7902
Adj R-Square	0.7876

(b) The model is significant with p-value less than .0001 and it can explain 79.02% of variation in log of median home value.

About the relationship between four predictors and median home value, the coefficients for age and indus are negative and estimated as -0.00247, -0.00624, respectively. There is a slight change in estimated coefficients after excluding three influential points. This result tells that the expected median of house value decreases with the multiplicative factors being 0.9975 and 0.9938, with one unit increase in age and indus, respectively. The coefficients for nox and rm are positive and estimated as 0.45177 and 0.35636, respectively. It means that the expected multiplicative increase in median of house value is 1.5711 and 1.4218, with one unit increase in nox and rm, respectively.

Model: MODEL1
Dependent Variable: logmedv

Analysis of Variance

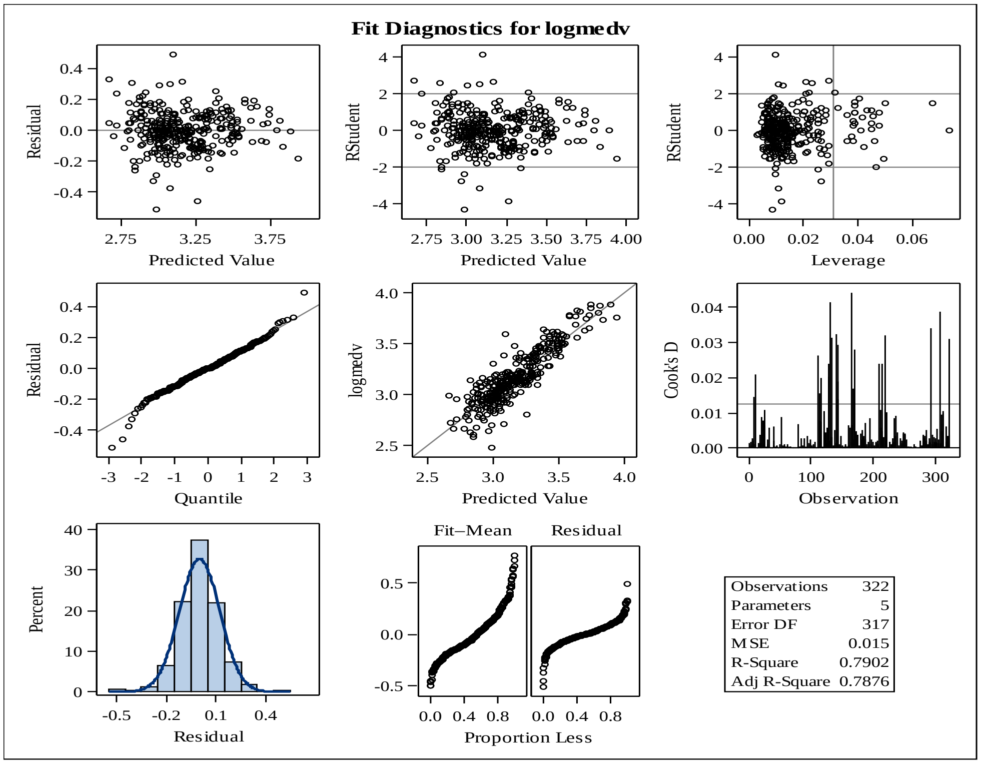
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	17.94849	4.48712	298.56	<.0001
Error	317	4.76429	0.01503		
Corrected Total	321	22.71278			

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

Parameter Estimates

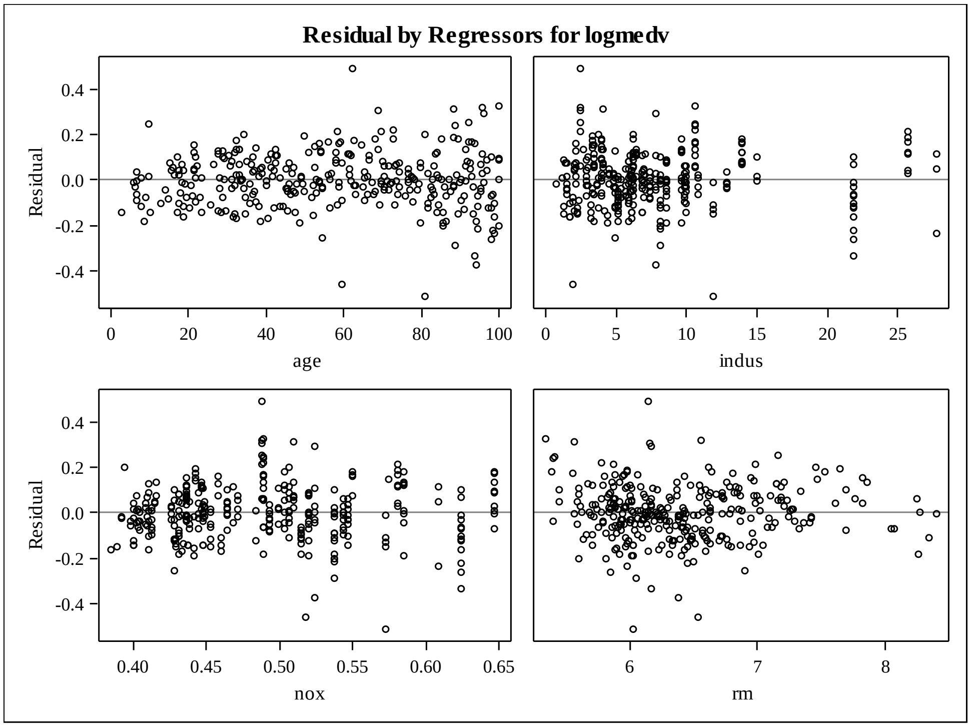
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.85758	0.10397	8.25	<.0001
age	1	-0.00247	0.00036252	-6.82	<.0001
indus	1	-0.00624	0.00163	-3.83	0.0002
nox	1	0.45177	0.16650	2.71	0.0070
rm	1	0.35636	0.01262	28.25	<.0001

Looking at the diagnostics, the residuals show no problematic trends as a function of predicted value and they mostly look normal. There are a couple of points that trail away from the line in the quantile plot, indicating a few potential outliers. The Cook's distances look reasonable, with none of the points being extremely large relative to the others. The residuals vs. predictor plots also look fine. While there are some minor deviations, such as a slight spreading on the right side of the age plot, there is no cause for concern about relationships between the predictor values and the residuals.



Fit Diagnostics for logmedv

Observations	322
Parameters	5
Error DF	317
MSE	0.015
R-Square	0.7902
Adj R-Square	0.7876



Residual by Regressors for logmedv