Exercise 1

(a) The data is not balanced, so the glm procedure is used.

Dependent Variable: logmedv

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	30.91593802	7.72898450	92.02	<.0001
Error	485	40.73787143	0.08399561		
Corrected Total	489	71.65380945			

R-Square	Coeff Var	Root MSE	logmedv Mean
0.431463	9.641833	0.289820	3.005859

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ptlevel	2	26.36639667	13.18319833	156.95	<.0001
over25kSqFt	1	1.89990854	1.89990854	22.62	<.0001
taxlevel	1	2.64963280	2.64963280	31.54	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ptlevel	2	2.82614807	1.41307404	16.82	<.0001
over25kSqFt	1	1.74134601	1.74134601	20.73	<.0001
taxlevel	1	2.64963280	2.64963280	31.54	<.0001

Firstly, p-values for all three terms, ptlevel, over25kSqFt and taxlevel, are less than .05, thus we can conclude that all three main effects are significant to explain the variation of log of median home values. In other words, the best main effects model includes all three terms. Specifically, the result from type I SS uses the amount of additional variation explained by the model when that term is added to the model containing the previous terms in the table, and results from type III SS are obtained from the amount of explained variation lost if we drop that term from the model. The overall model is significant with p-value less than .001 and the amount of variation in log median home value explained by the model is 43.15%. Again, all three individual terms in the model are significant with p-value less than .05.

(b) To examine the mean differences in expected log median home values across groups, least squares means are compared and the confidence intervals for differences are derived as follows.

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

ptlevel	logmedv LSMEAN	LSMEAN Number
higher	2.83889867	1
lower	3.09610642	2
medium	3.07921066	3

Least Squares Means for Effect ptlevel				
i	**j**	**Difference Between Means**	**Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)**	
1	**2**	-0.257208	-0.390760	-0.123655
1	**3**	-0.240312	-0.340386	-0.140238
2	**3**	0.016896	-0.089799	0.123591

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

over25kSqFt	logmedv LSMEAN	H0:LSMean1=LSMean2
		Pr > |t|
none	2.92842993	<.0001
some	3.08104724	

Least Squares Means for Effect over25kSqFt				
i	**j**	**Difference Between Means**	**Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)**	
1	**2**	-0.152617	-0.218477	-0.086757

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

taxlevel	logmedv LSMEAN	H0:LSMean1=LSMean2
		Pr > |t|
higher	2.88069085	<.0001
lower	3.12878632	

Least Squares Means for Effect taxlevel				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.248095	-0.334889	-0.161302

Firstly, with respect to ptlevel, areas with higher pupil-teacher ratio have significantly lower expected log median home values than areas with lower or medium pupil-teacher ratio, and the estimated differences between (ptlevel1 - ptlevel2) and (ptlevel1- ptlevel3) are -0.26 and -0.24. However, the two groups with lower pupil-teacher ratio and medium pupil-teacher ratio are not significantly different in the aspect of expected log median home values. Secondly, tables for over25kSqFt shows that if there is residential land zoned for lots over 25,000 square feet, the expected log median home values are significantly higher than areas without residential land zoned for lots over 25,000 square feet. The estimated differences between (none-some) is -0.15 and the confidence interval does not contain zero. Lastly, in terms of taxlevel, areas with lower tax level have significantly higher expected log median home values than areas with higher tax level. The estimated difference between two groups is -0.25 and the confidence interval does not contain zero.

Exercise 2

Dependent Variable: logmedv

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ptlevel	2	26.36639667	13.18319833	168.21	<.0001
over25kSqFt	1	1.89990854	1.89990854	24.24	<.0001
over25kSqFt*ptlevel	2	2.87861174	1.43930587	18.36	<.0001
taxlevel	1	2.65412951	2.65412951	33.86	<.0001
ptlevel*taxlevel	0	0.00000000	.	.	.
over25kSqFt*taxlevel	0	0.00000000	.	.	.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ptlevel	2	0.91713735	0.45856868	5.85	0.0031
over25kSqFt	1	1.18090996	1.18090996	15.07	0.0001
over25kSqFt*ptlevel	2	2.88310845	1.44155423	18.39	<.0001
taxlevel	1	2.65412951	2.65412951	33.86	<.0001
ptlevel*taxlevel	0	0.00000000	.	.	.
over25kSqFt*taxlevel	0	0.00000000	.	.	.

(a) Here we start by adding all of the two way interaction to the model. Both the results from type I SS and type III SS tell us all of the three main effects and only the interaction between over25kSqFt and ptlevel are significant. The other two interaction terms provide no additional explained variation at all. So we refit an anova model with the selected terms. The results are shown as follows.

Dependent Variable: logmedv

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	33.79904647	5.63317441	71.88	<.0001
Error	483	37.85476298	0.07837425		
Corrected Total	489	71.65380945			

R-Square	Coeff Var	Root MSE	logmedv Mean
0.471699	9.313610	0.279954	3.005859

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ptlevel	2	26.36639667	13.18319833	168.21	<.0001
over25kSqFt	1	1.89990854	1.89990854	24.24	<.0001
taxlevel	1	2.64963280	2.64963280	33.81	<.0001
over25kSqFt*ptlevel	2	2.88310845	1.44155423	18.39	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ptlevel	2	0.91713735	0.45856868	5.85	0.0031
over25kSqFt	1	1.18090996	1.18090996	15.07	0.0001
taxlevel	1	2.65412951	2.65412951	33.86	<.0001
over25kSqFt*ptlevel	2	2.88310845	1.44155423	18.39	<.0001

The overall model is significant with p-value less than .001 and the amount of variation explained by the model is 47.17%. All three individual terms and the interaction term in the model are significant with p-value less than .05.

(b) Again, we can determine the significant differences among groups by investigating least squares means.

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

ptlevel	logmedv LSMEAN	LSMEAN Number
higher	2.79691691	1
lower	3.10765222	2
medium	3.07024758	3

Least Squares Means for Effect ptlevel				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.310735	-0.526945	-0.094525
1	3	-0.273331	-0.472547	-0.074114
2	3	0.037405	-0.065976	0.140786

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

over25kSqFt	logmedv LSMEAN	H0:LSMean1=LSMean2
		Pr > |t|
none	2.87032298	0.0001
some	3.11288816	

Least Squares Means for Effect over25kSqFt				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.242565	-0.365350	-0.119780

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

taxlevel	logmedv LSMEAN	H0:LSMean1=LSMean2
		Pr > |t|
higher	2.86549316	<.0001
lower	3.11771798	

Least Squares Means for Effect taxlevel				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.252225	-0.337388	-0.167062

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

over25kSqFt	ptlevel	logmedv LSMEAN	LSMEAN Number
none	higher	2.76465470	1
none	lower	2.80903925	2
none	medium	3.03727500	3
some	higher	2.82917912	4
some	lower	3.40626518	5
some	medium	3.10322017	6

Least Squares Means for Effect over25kSqFt*ptlevel				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.044385	-0.231989	0.143220
1	3	-0.272620	-0.395244	-0.149997
1	4	-0.064524	-0.538220	0.409171
1	5	-0.641610	-0.837606	-0.445615
1	6	-0.338565	-0.467977	-0.209154
2	3	-0.228236	-0.399103	-0.057369
2	4	-0.020140	-0.508552	0.468272
2	5	-0.597226	-0.826514	-0.367937
2	6	-0.294181	-0.469983	-0.118379
3	4	0.208096	-0.259224	0.675415
3	5	-0.368990	-0.549030	-0.188950
3	6	-0.065945	-0.169616	0.037725
4	5	-0.577086	-1.068782	-0.085390
4	6	-0.274041	-0.743187	0.195105
5	6	0.303045	0.118315	0.487775

Firstly, with respect to ptlevel, areas with higher pupil-teacher ratio have significantly lower expected log median home values than areas with lower or medium pupil-teacher ratio, and the estimated differences between (ptlevel1 - ptlevel2) and (ptlevel1- ptlevel3) are -0.31 and -0.27. Similar to exercise 1, two groups with lower pupil-teacher ratio and medium pupil-teacher ratio are not significantly different in the aspect of expected log median home values. Secondly, tables for over25kSqFt show that if there is residential land zoned for lots over 25,000 square feet, the expected log median home values are significantly higher than areas without residential land zoned for lots over 25,000 square feet. The estimated differences between (none-some) is -0.24 and the confidence interval does not contain zero. In terms of taxlevel, areas with lower tax level have significantly higher expected log median home values than areas with higher tax level. The estimated difference between the two groups is -0.25 and confidence interval does not contain zero. The relationships are similar to what we saw in exercise 1, but the magnitudes of the differences are slightly different.

Lastly, when we see the results from the interaction term, if there is no residential land zoned for lots over 25,000 square feet, the expected log median home values don't have significant difference between the lower and higher pupil-teacher ratio groups. If there is no residential land zoned for lots over 25,000 square feet, expected home values are higher in the medium pupil-teacher areas than in the lower or higher areas.

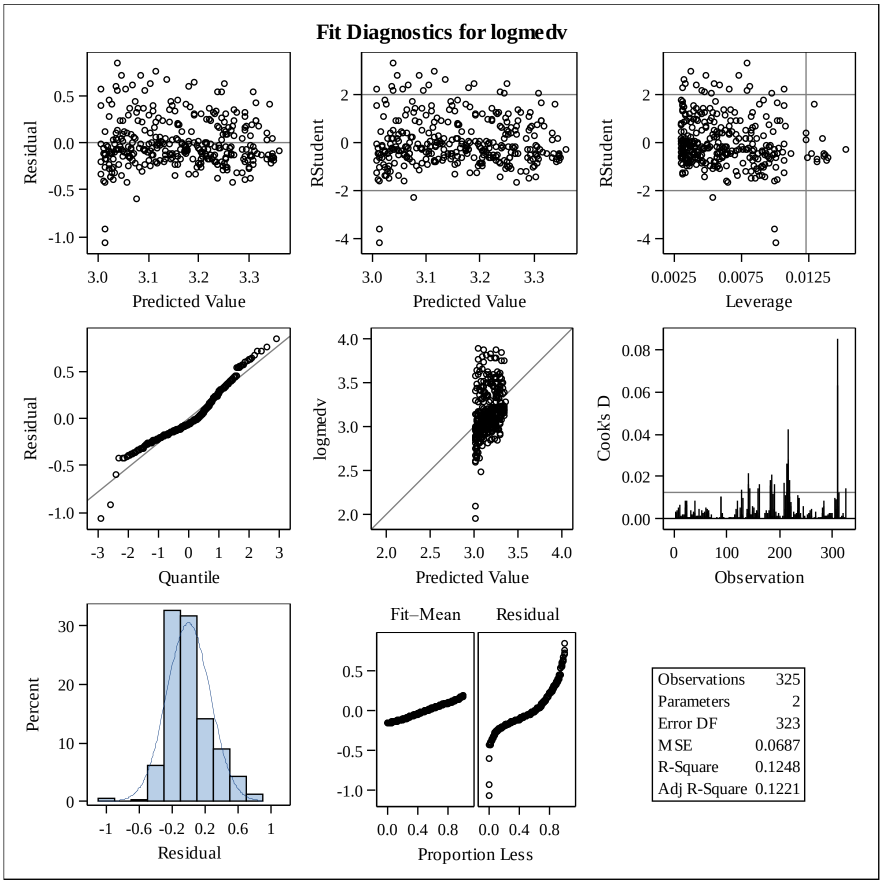
However, if there is residential land zoned for lots over 25,000 square feet, the expected log median home values are significantly higher when the pupil-teacher ratio is lower, compared to higher or medium pupil-teacher ratio groups. We can clearly see that for each over25kSqFt group, the expected log median home values have different behavior for lower pupil-teacher ratio. It was not captured in exercise 1, when only main effects are used in the model.

When comparing the none and some groups interacting with pupil-teacher ratios, we see that numerous areas with no residential land zoned for lots over 25,000 square feet have significantly lower log median home values. Specifically, none with higher pupil-teacher ratio has lower values than some with lower or medium pupil-teacher ratios, none with lower pupil-teacher ratios has lower values than some with medium pupil-teacher ratios, and none with medium pupil-teacher ratios has lower values than some with lower pupil-teacher ratios.

Exercise 3

(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. Some observations show standardized residuals greater than 2, but it does not look serious. If those points have undue influence, we will remove them based when we look at Cook's distances. Also no pattern is detected in the residual 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 value compared to others, so we will delete points in the model that have Cook's distance greater than 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.

Model: MODEL1
Dependent Variable: logmedv



(b) The 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 for a one year increase in house age, the expected median of house value is multiplied by exp(-0.00328)=0.9967 indicating a slight multiplicative decrease. About the diagnostic plots, since we delete some potential outliers, now residual plot and Cook's distance plot look better than the results from part (a). The normal QQ plot still shows some points not lying in the straight line, but it seems not serious. However, 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.

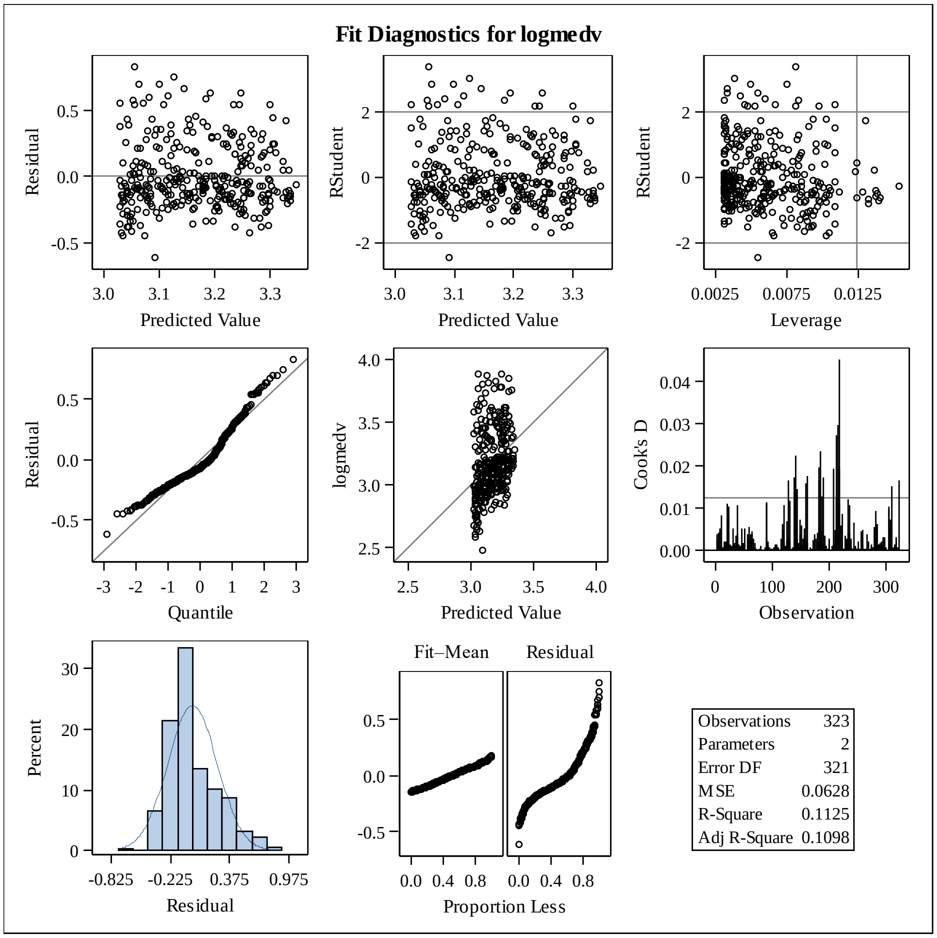
Model: MODEL1
Dependent Variable: logmedv

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

Model: MODEL1
Dependent Variable: logmedv



Observations	323
Parameters	2
Error DF	321
MSE	0.0628
R-Square	0.1125
Adj R-Square	0.1098

Exercise 4

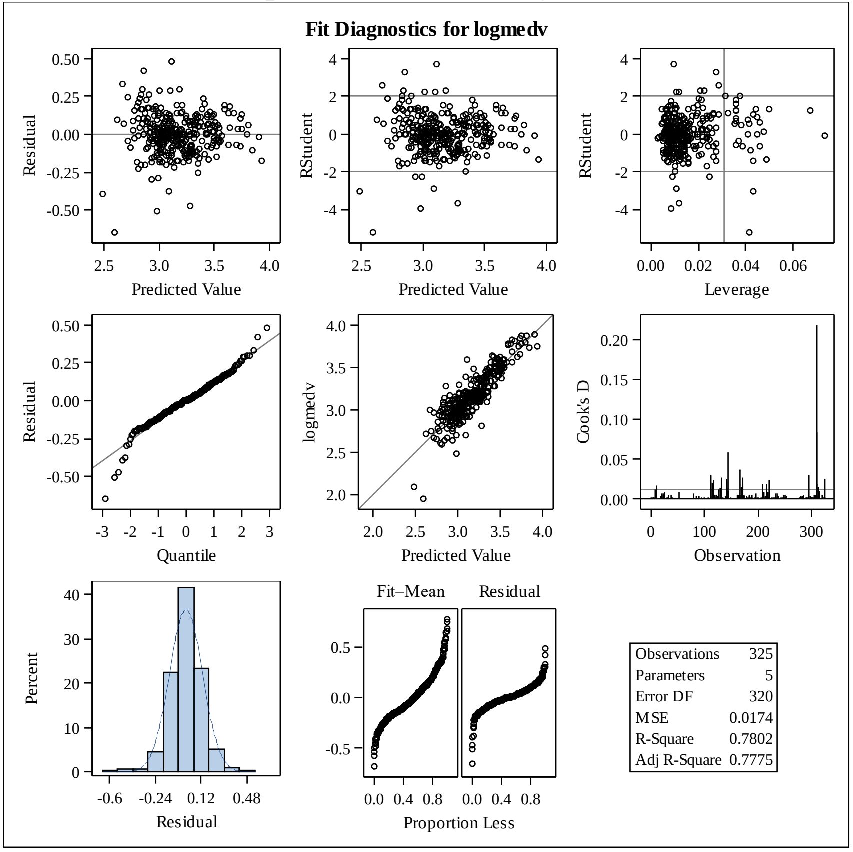
(a) The summary of stepwise selection is shown in the table below. It suggests that all of the four variables are significant and should be kept in the model.

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

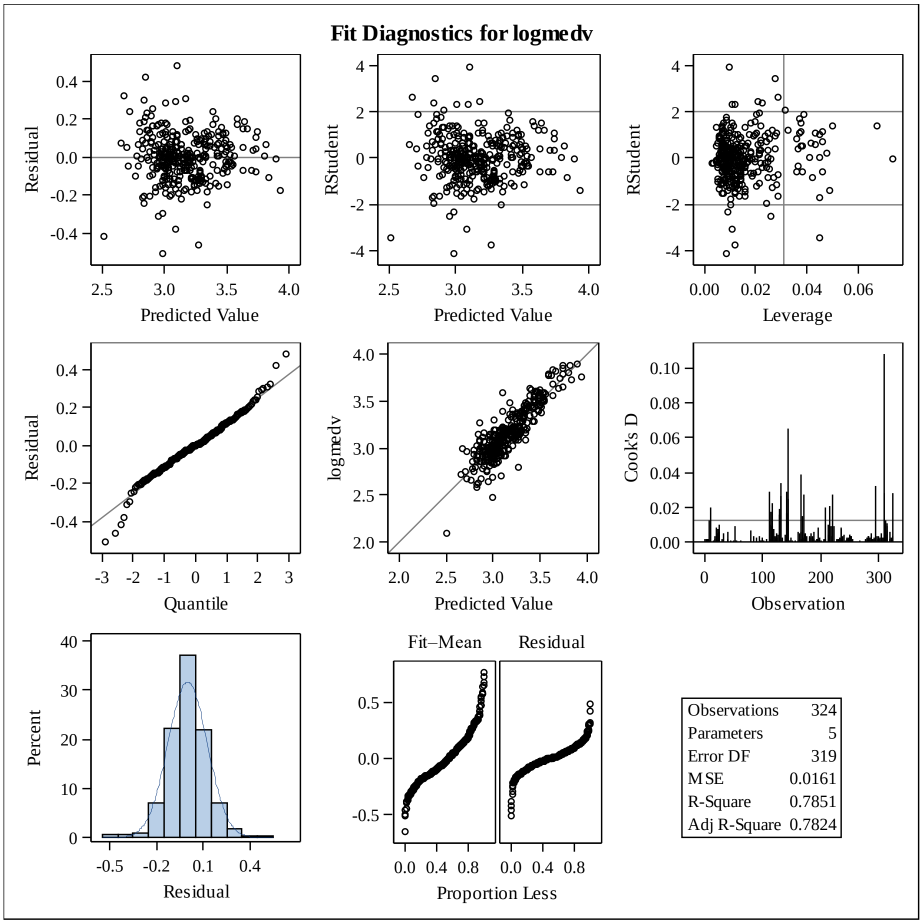
The following are the diagnostic plots. There is an observation with standardized residual below -4, which may need a careful look. No pattern is detected in the residual plot. For QQ plot, some points are not lying in the straight line, but the histogram from residuals looks very symmetric and bell-shaped, thus normality assumption for error term looks valid. 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.

Model: MODEL1
Dependent Variable: logmedv



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 0.015*4=0.06. Then the same model is re-fitted after deleting those potentially unduly influential points.

Model: MODEL1
Dependent Variable: logmedv



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. The coefficients for age and indus are negative and estimated as -0.00247, -0.00624 respectively. It means that if one unit increase in age or indus will lead to a decrease of the expected median of house value with, with the multiplicative factors being 0.9975 and 0.9938, respectively. The coefficients for nox and rm are positive and estimated as 0.45177 and 0.35636, respectively. It means that if one unit increase in nox will lead to an expected multiplicative increase of 1.5711 and one additional room on average would lead to an expected multiplicative increase of 1.4281.

Model: MODEL1
Dependent Variable: logmedv

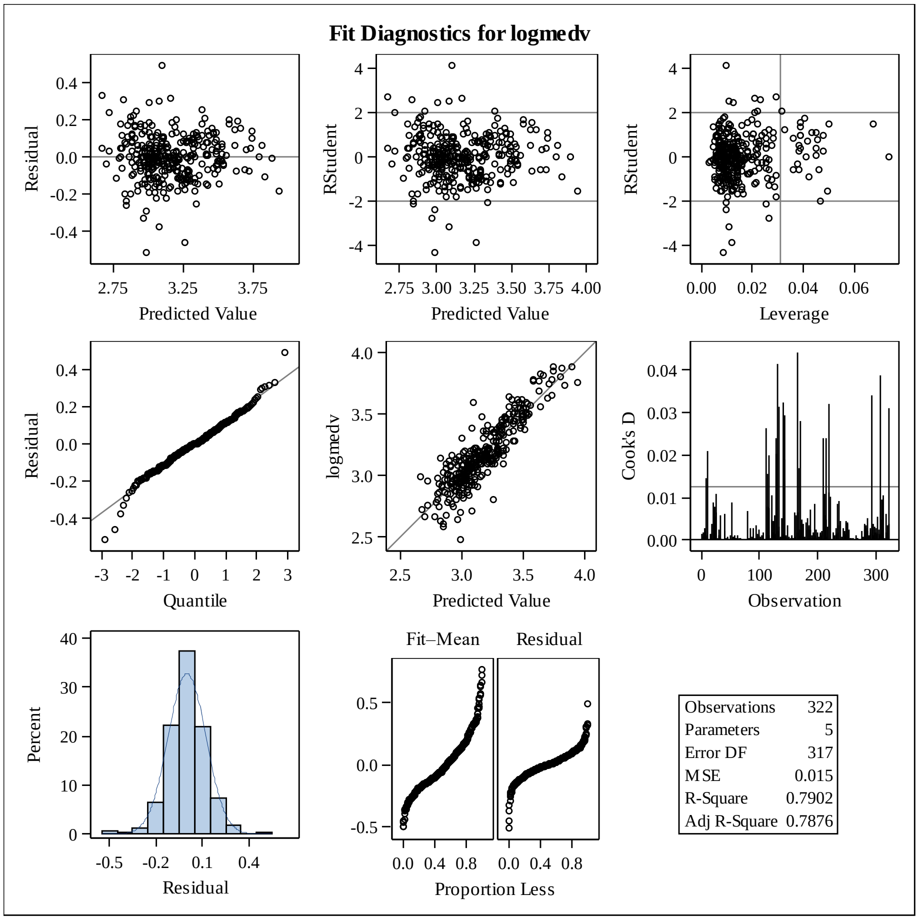
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	17.94849	4.48712	298.56	<.0001
Error	317	4.76429	0.01503		
Corrected Total	321	22.71278			

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

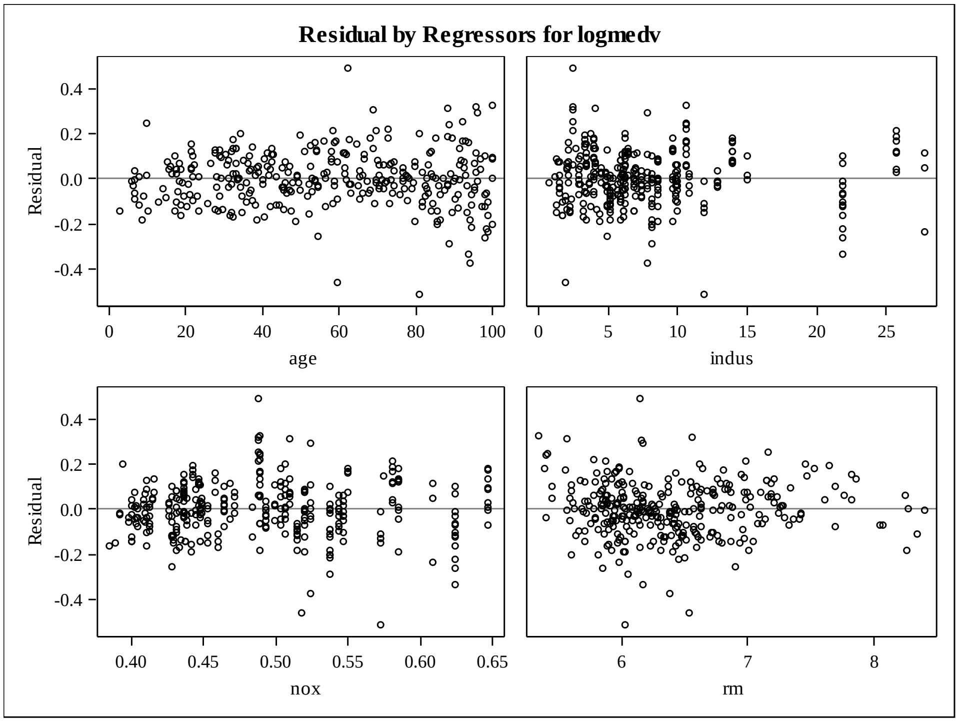
About the diagnostic plots, after we delete some potential outliers, the residual plot and Cook’s distance plot look better than the results from part (a). The residual vs predictor plots don’t show any pattern. The normal QQ plot still shows some points not lying in the straight line, but it seems not serious.

Model: MODEL1
Dependent Variable: logmedv



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