STAT 448 HW3
Jinran Yang

Problem 1 a

Cylinders	Origin	Type	MPG (Highway) Mean	Std	N
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

Because the counts of each cell are different, this data is unbalanced.

Apparent differences are as follow:

1. the mean value of MPG_Highway of car with 6 cylinders generally less than with 4 cylinders, which means cars with 6 cylinders tend to have lower fuel efficiency compared to those with 4 cylinders.
2. The standard deviation of cars with 6 cylinders are smaller than those with 4 cars, which imply that the fuel efficiency of cars with 6 cylinders are more uniform.
3. When it comes to the cars with 4 cylinders and origin in Asia, type in Sports have lower fuel efficiency than type in Sedan. Since it has a smaller MPG_Highvalue.
4. Among cars with 6 cylinders, cars whose origin are Asia have lower standard deviation compared to those origin are USA. So, the fuel efficiency of cars whose origin is Asia tend to more uniform.
5. Cars type in Sports and origin is Asia has the lowest fuel efficiency among any other cars.

And cars type in Sedan and origin is Asia has the highest fuel efficiency among any other cars.

Problem 1 b

(1) Start with a three-way main effects ANOVA.

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

With unbalanced data, we should use ***proc glm***.

Begin with the model included all three main effects, according to the P value in the first table is less than .0001, we can know the model is significant at first. Then by looking at the P value of each main effect, we can see that the P value of origin is much larger than .05, so the origin is no significant. This variable should not be keep in model. And the P value of the cylinders and type are both smaller than .05, so these two variables are significant and we should keep them.

（2）Building a model with cylinders and type

The GLM Procedure

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

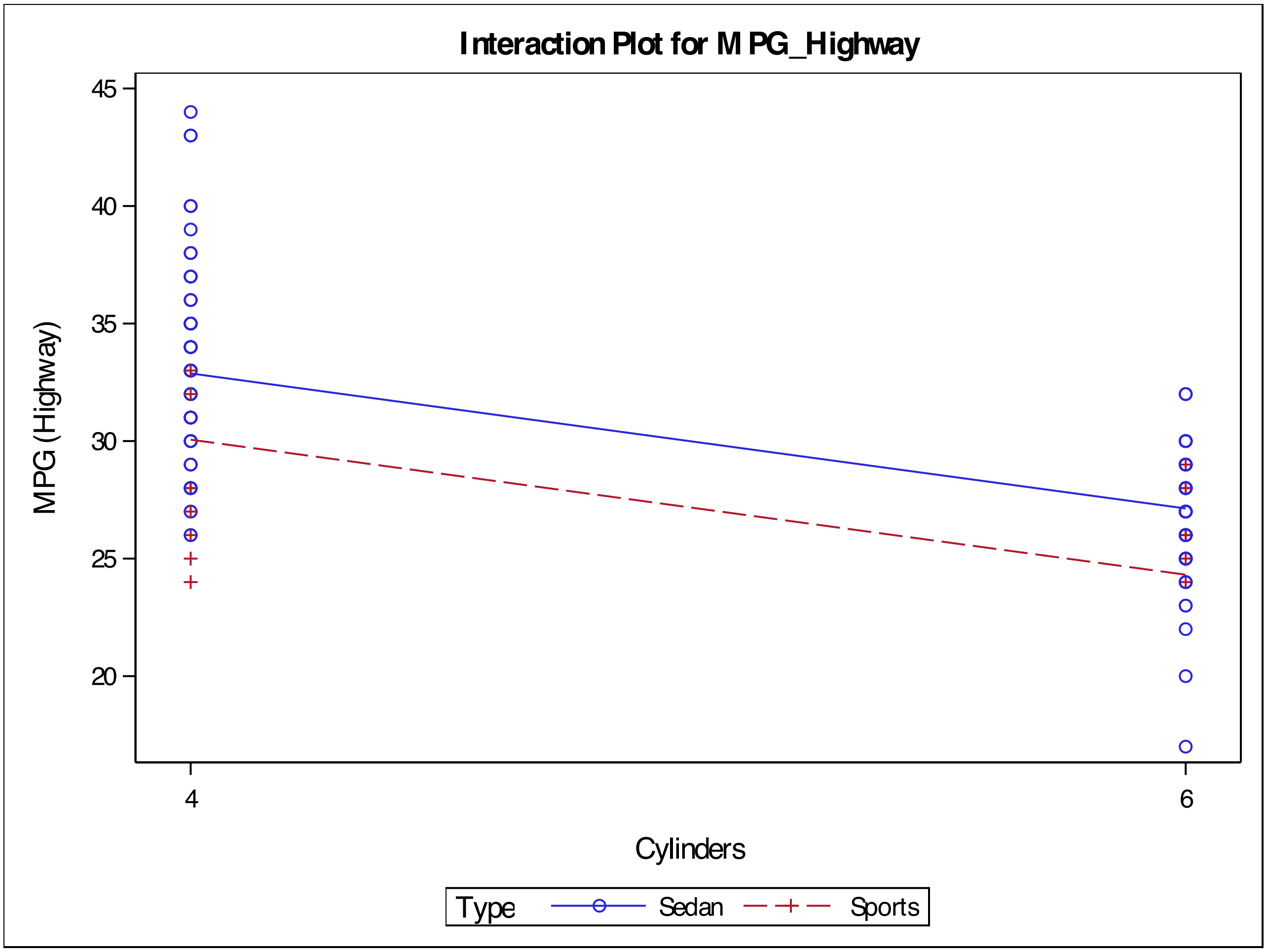
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cylinders	1	1470.787732	1470.787732	138.22	<.0001
Type	1	115.780611	115.780611	10.88	0.0012

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 P value in the first table is less than .0001, we can know the model is significant at first. And both two categorical variables are significant, since their P value are all smaller than .05. And the R square value in this module is the same as the previous, which prove that origin is not significant in this module.

This model describes 45.7% variation in highway fuel efficiency (R square equal to 0.457).

The GLM Procedure



Problem 1 c

Add type*Cylinders to the model:

The GLM Procedure

Dependent Variable: MPG_Highway MPG (Highway)

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 P value in the first table is less than .0001, we can know the model is significant at first. Then by looking at the P value of each effect, we know that they are all significant. So, we should include type*cylinders in our module. This module describes 48.14% variation in highway fuel efficiency (R square equal to 0.4814).

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

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

As we can see from the table, the mean value of MPG_High of cars type in Sedan is higher than that of cars type in Sports, which means Sedan cars have higher fuel efficiency in general. The P value of test is smaller than 0.05 and the 95% Confidence Limits doesn’t contain 0, thus the test result is valid.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

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

As we can see from the table, the mean value of MPG_High of cars with 4 Cylinders is higher cars with 6 Cylinders, which means 4 Cylinders cars have higher fuel efficiency in general. The P value of test is smaller than 0.05 and the 95% Confidence Limits doesn’t contain 0, thus the test result is valid.

The GLM Procedure
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

As we can see from the table,

(1) mean difference between Sedan cars with 4 Cylinders and Sports cars with 4 Cylinders is significant (8.3). The P value of test is smaller than 0.05 and the 95% Confidence Limits doesn’t contain 0, thus the test result is valid.

(2) mean difference between Sedan cars with 4 Cylinders and Sedan cars with 6 Cylinders is significant (7.5). The P value of test is smaller than 0.05 and the 95% Confidence Limits doesn’t contain 0, thus the test result is valid.

(3) mean difference between Sedan cars with 4 Cylinders and Sports n cars with 6 Cylinders is significant (9.7). The P value of test is smaller than 0.05 and the 95% Confidence Limits doesn’t contain 0, thus the test result is valid.

(4) Any other differences of interaction groups are not significant.

Problem 2 a

The REG Procedure
Model: MODEL1
Dependent Variable: logmedv

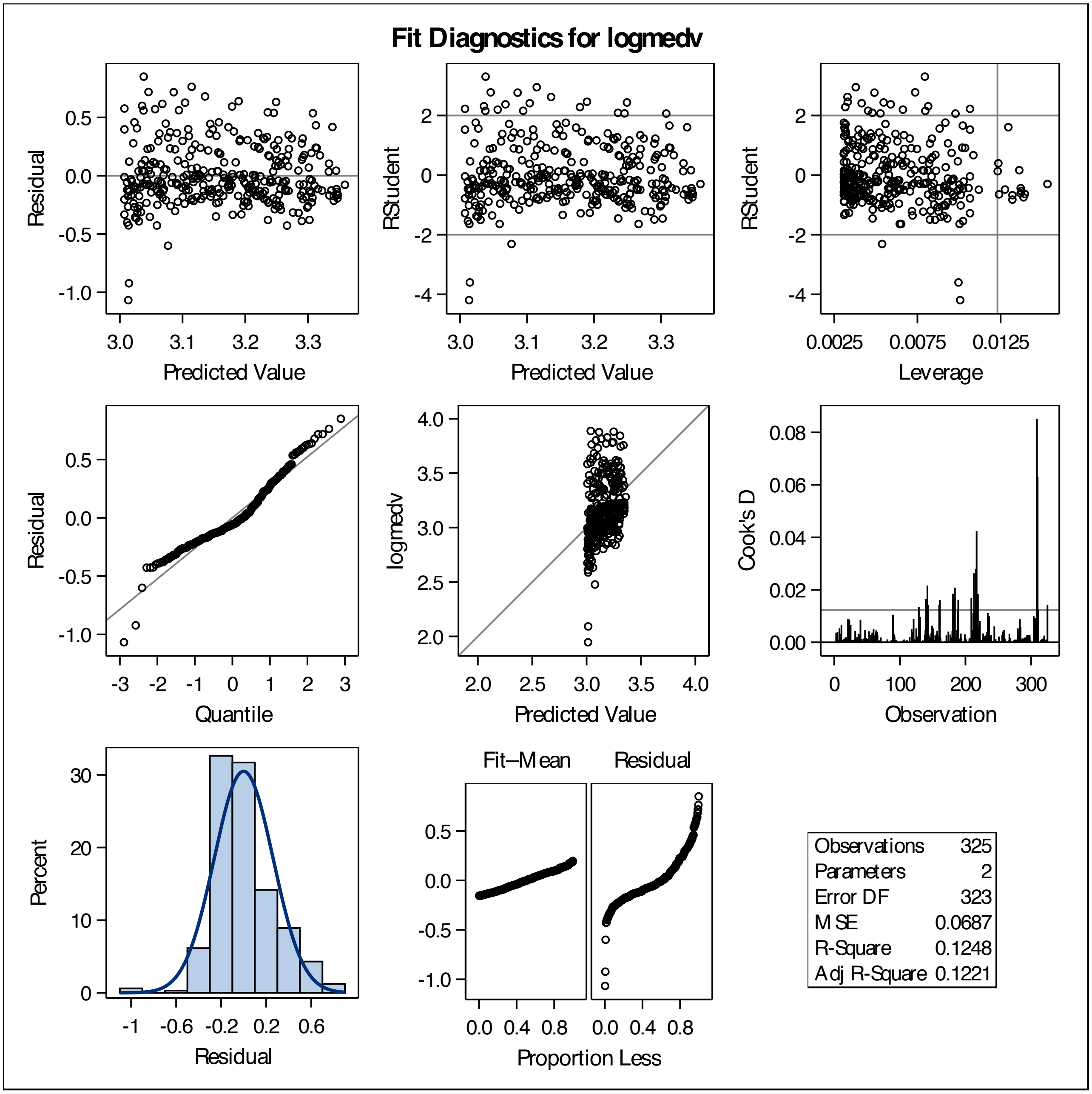
Number of Observations Read	325
Number of Observations Used	325

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			

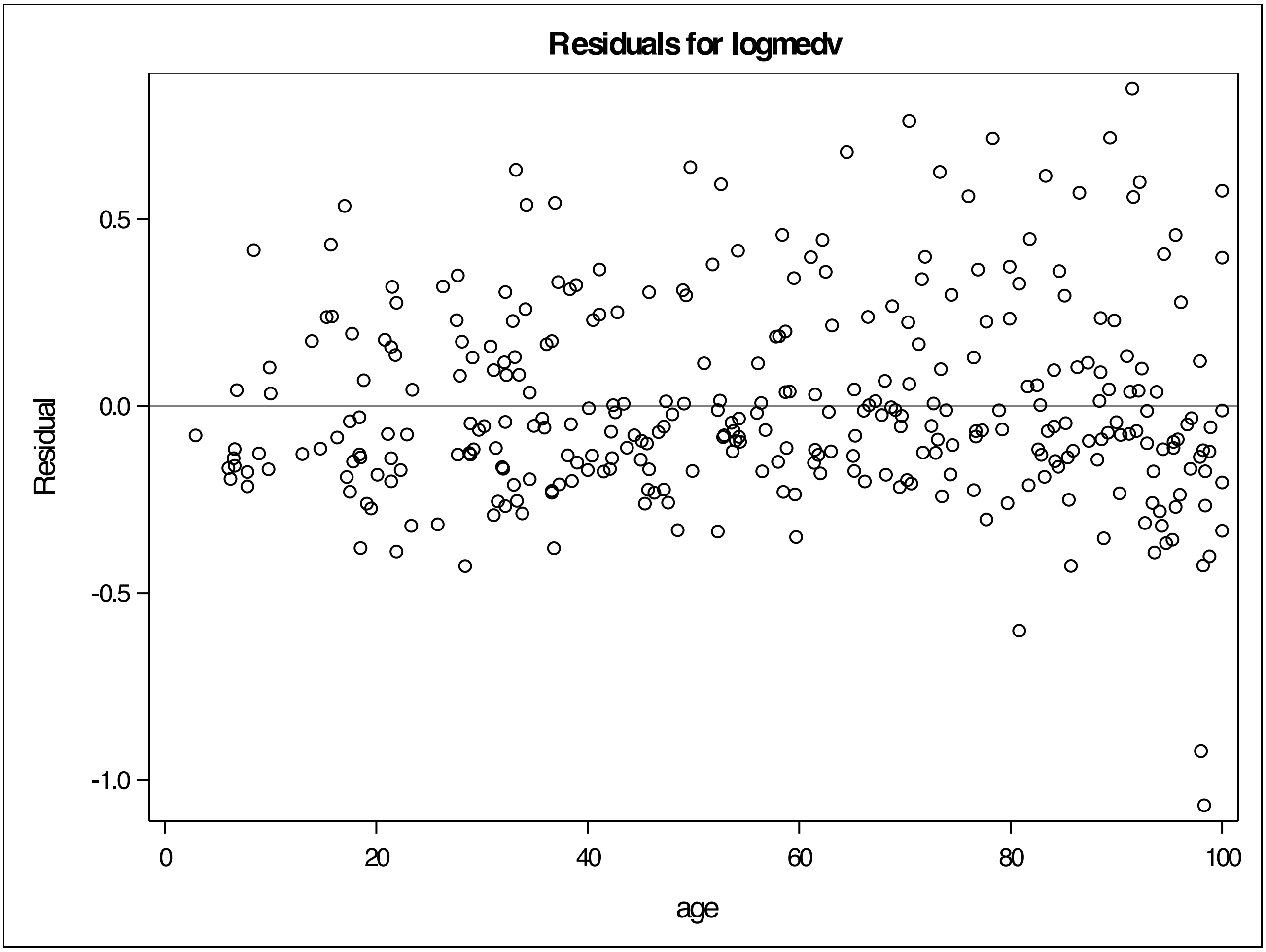
Root MSE	0.26216	**R-Square**	0.1248
Dependent Mean	3.16225	**Adj R-Sq**	0.1221
Coeff Var	8.29017		

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

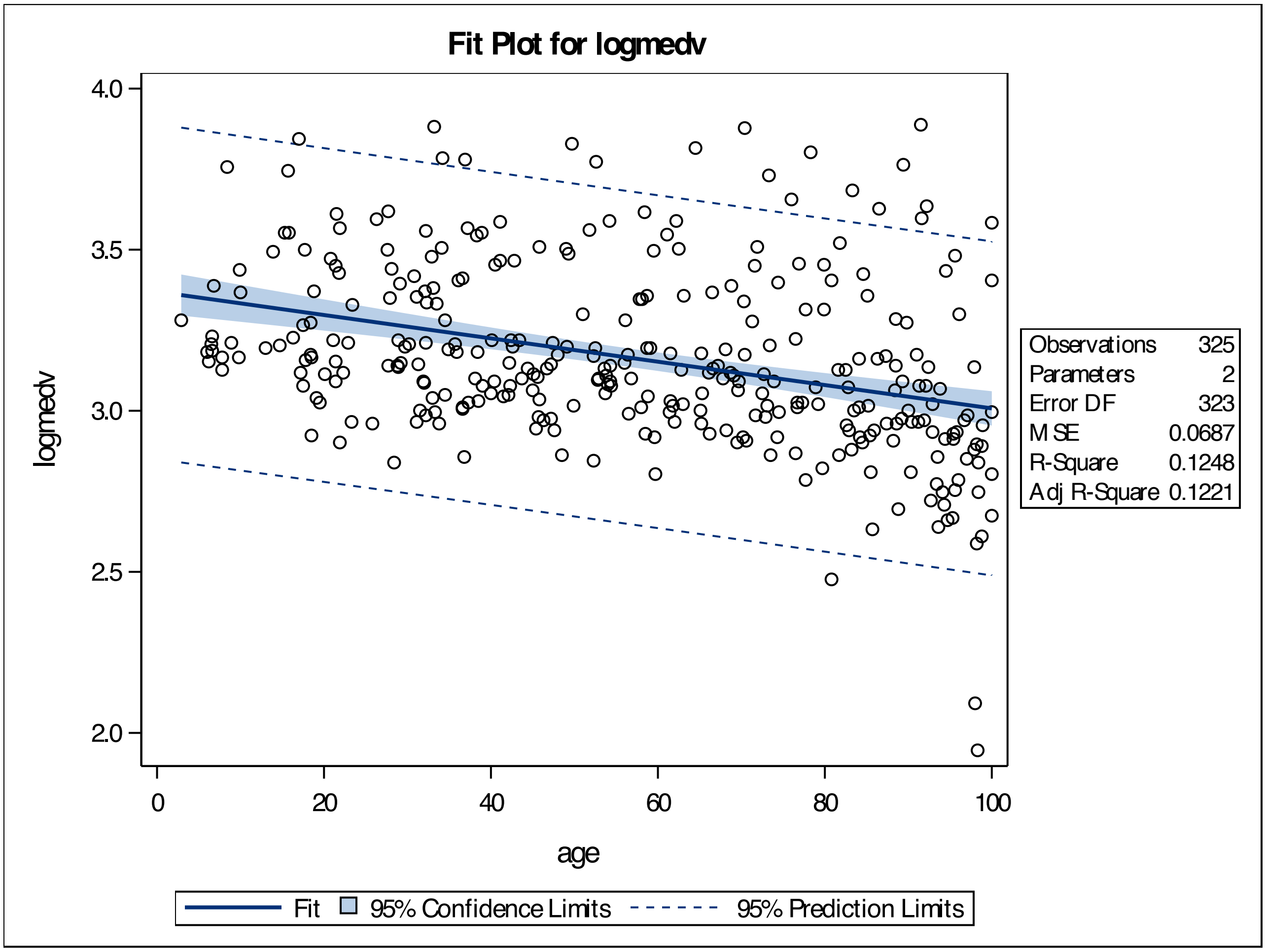
The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



By looking at the cook's difference, we can find that there exists some influential point in this data set, we should exclude them.

The cut-off cook's difference is 4/N approximately, N is the number of observations in the data set.
Thus, the cut-off line in this data set is 4/325= 0.012 approximately.
According to the context we should remove the data with cd larger than 0.048.

The REG Procedure
Model: MODEL1
Dependent Variable: logmedv

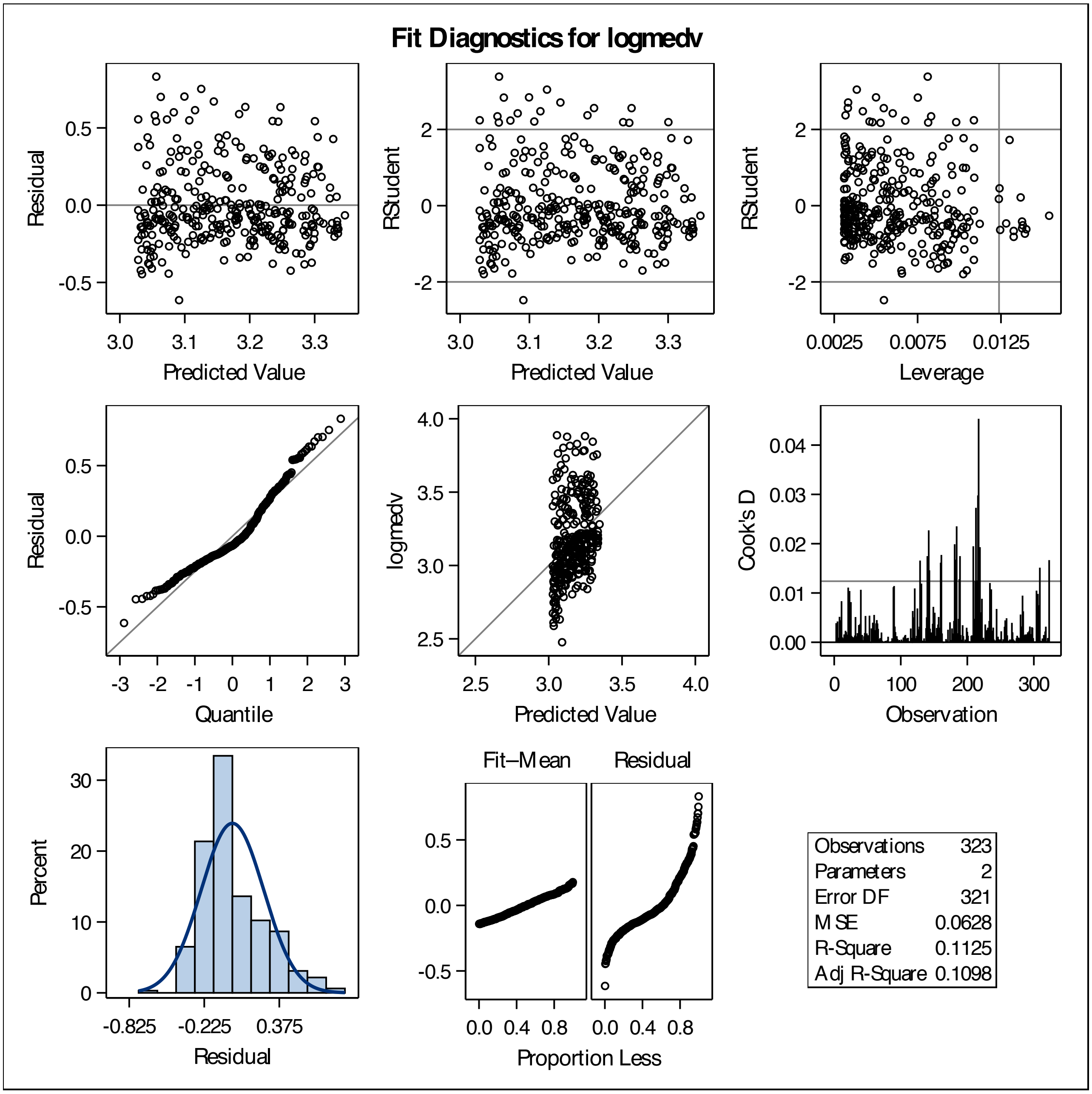
Number of Observations Read	323
Number of Observations Used	323

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			

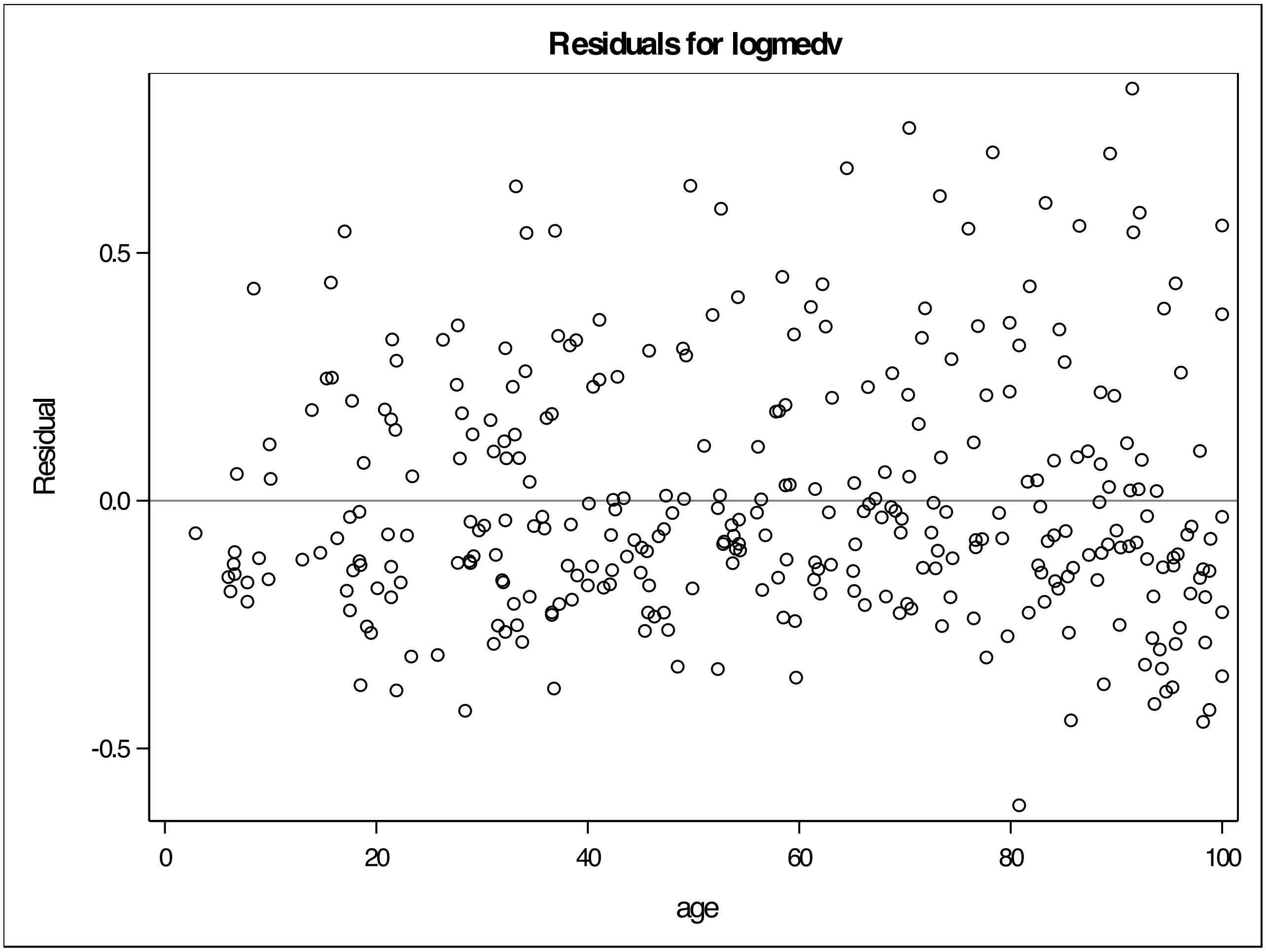
Root MSE	0.25065	**R-Square**	0.1125
Dependent Mean	3.16933	**Adj R-Sq**	0.1098
Coeff Var	7.90859		

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

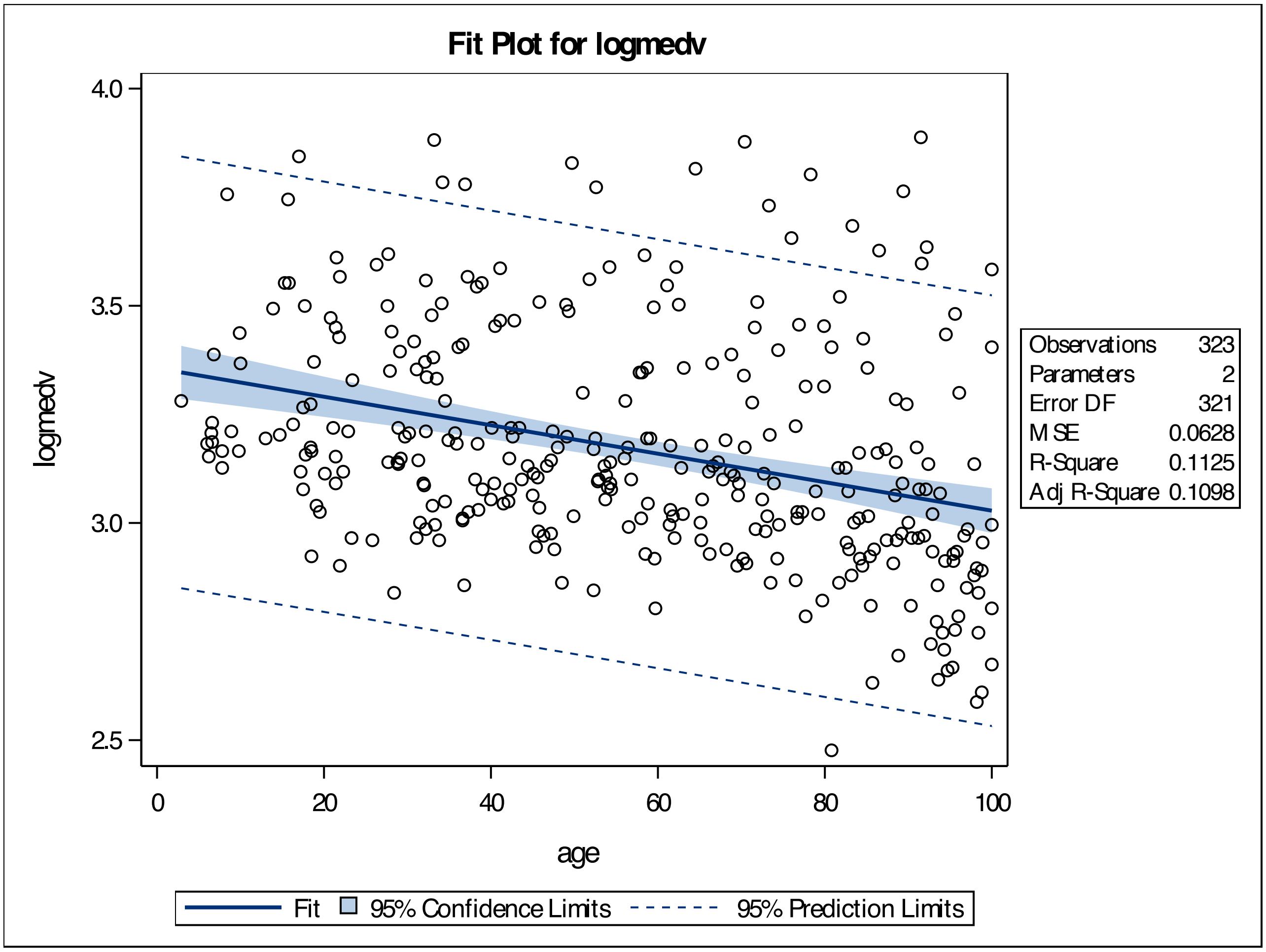
The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



Problem 2 b

We can tell from the cook's distance graph that influential points have been removed.

The **Parameter Estimate** of age and the log of median home value is -0.00328. As a result, one unit of age increase will lead to the median home value become e^(-0.00328) multiply the original median value. Since e^(-0.00328) is less than 1, the median value is actually decrease. So, there are negative relationship between the home age and the home median value.

The model describes 11.25% variation in log of median home value.

According to the plot of residual, we can easily find that the residuals are not normally distributed—some points in Rstudent graph lie outside -2 and 2, histogram is right-skewed, Q-Q plot is not a straight line. So, we might need other predictors.

The REG Procedure
Model: MODEL1
Dependent Variable: logmedv

As far as I am concerned, this model with only one predictor--age is not very useful. First, the value of R square is small (0.1125) which means it can just explain 11.25% variation and the residuals are not normally distributed, so I think this model should be further improved.

Problem 3 a

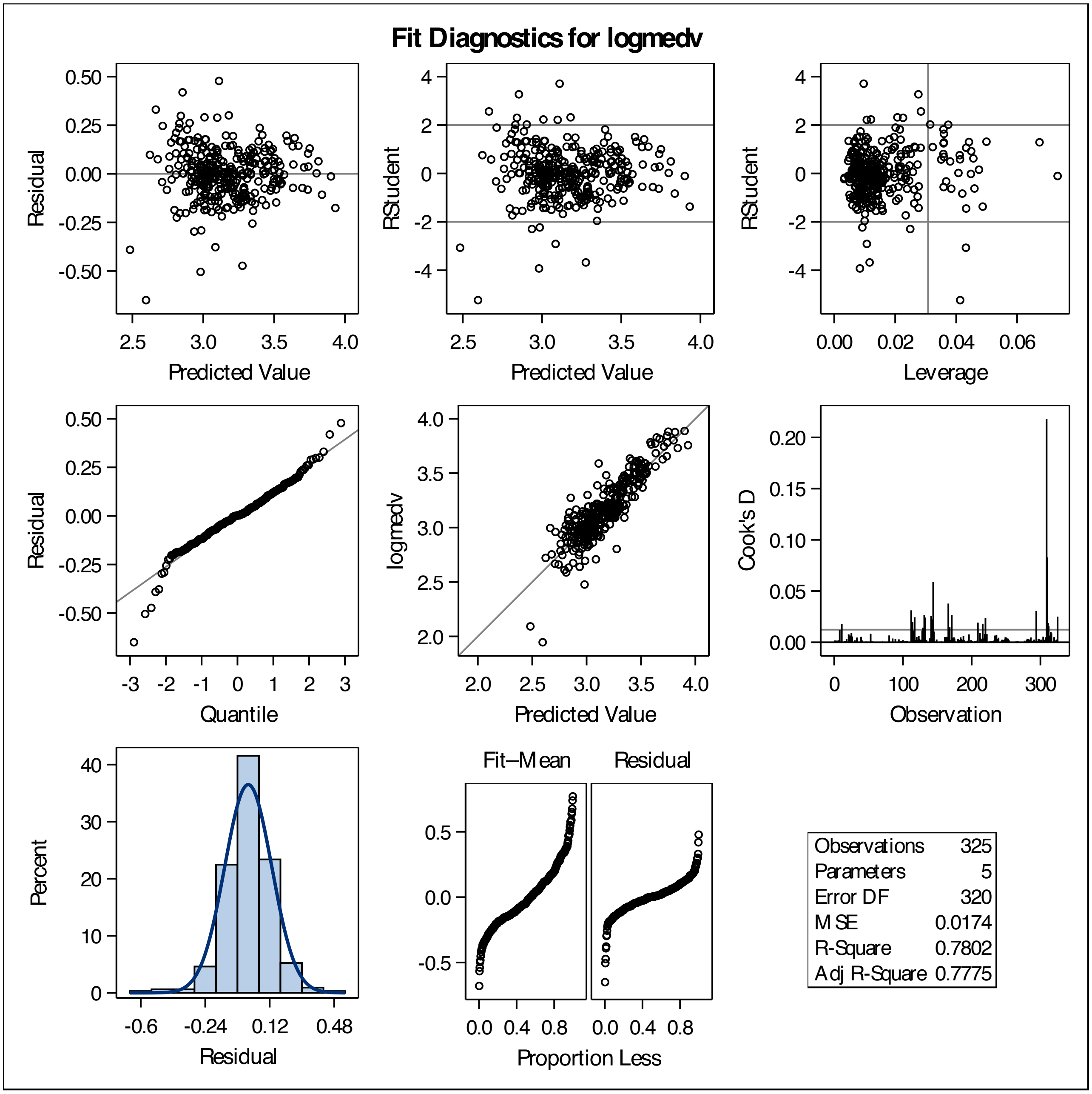
After checking the count of each cell in cross-tabulation, we know that this data is a balanced data set.

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|**
Intercept	1	0.87577	0.11130	7.87	<.0001
age	1	-0.00238	0.00038830	-6.13	<.0001
indus	1	-0.00896	0.00171	-5.24	<.0001
nox	1	0.46342	0.17909	2.59	0.0101
rm	1	0.35470	0.01349	26.29	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



By looking at the cook's difference, we can find that there exists some influential point in this data set, we should exclude them.

According to the plot of residual, we can easily find that the residuals are normal distributed—majority of points in Rstudent graph lie between -2 and 2, histogram is symmetric, Q-Q plot is almost a straight line. So, this model is reasonable.

Problem 3 b

The CORR Procedure

Check the multicollinearity:

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
age	325	57.23262	27.28745	18601	2.90000	100.00000
indus	325	7.72529	5.73225	2511	0.74000	27.74000
nox	325	0.49053	0.06577	159.42360	0.38500	0.64700
rm	325	6.38409	0.59478	2075	5.09300	8.39800

Pearson Correlation Coefficients, N = 325 Prob > |r| under H0: Rho=0				
	age	**indus**	**nox**	**rm**
age	1.00000	0.44917 <.0001	0.72134 <.0001	-0.15587 0.0049
indus	0.44917 <.0001	1.00000	0.59326 <.0001	-0.40195 <.0001
nox	0.72134 <.0001	0.59326 <.0001	1.00000	-0.19129 0.0005
rm	-0.15587 0.0049	-0.40195 <.0001	-0.19129 0.0005	1.00000

By looking at the **Pearson Correlations**, we know that there are not any variable perfectly correlated to others.

The REG Procedure
Model: MODEL1
Dependent Variable: logmedv

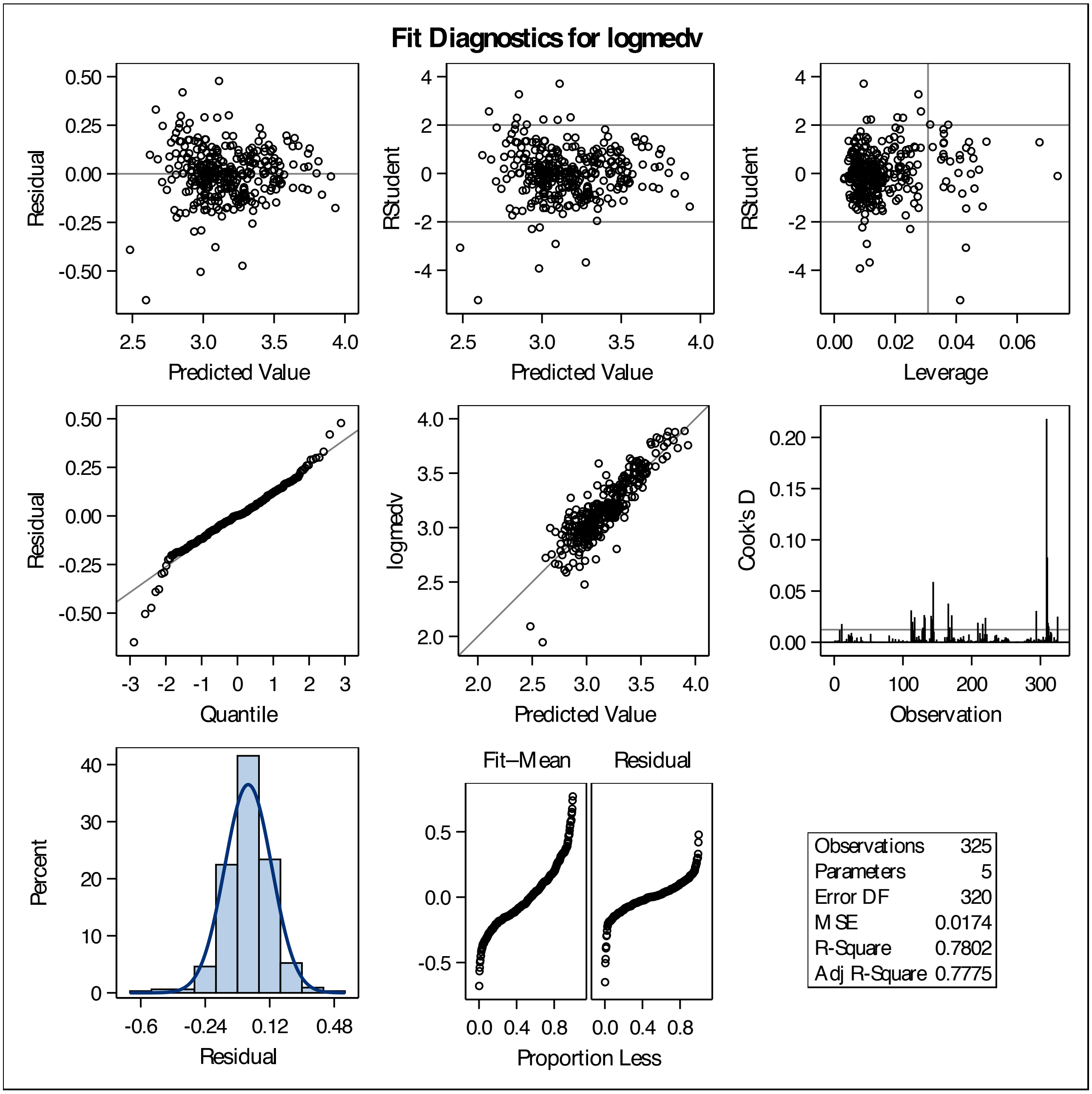
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	19.79088	4.94772	284.04	<.0001
Error	320	5.57415	0.01742		
Corrected Total	324	25.36503			

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

By looking at the value of **Variance Inflations**, we know that there are nor any variable highly correlated to others. Since none of the variable’s VIF larger than 10.

The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



As far as I am concerned, this model an ok model. It has high value of R square which means this model explains 78% variance; All predictors are not highly or perfectly correlative to others; Residuals are normal distributed.
But it has some influential points which should be excluded.

Problem 4 a

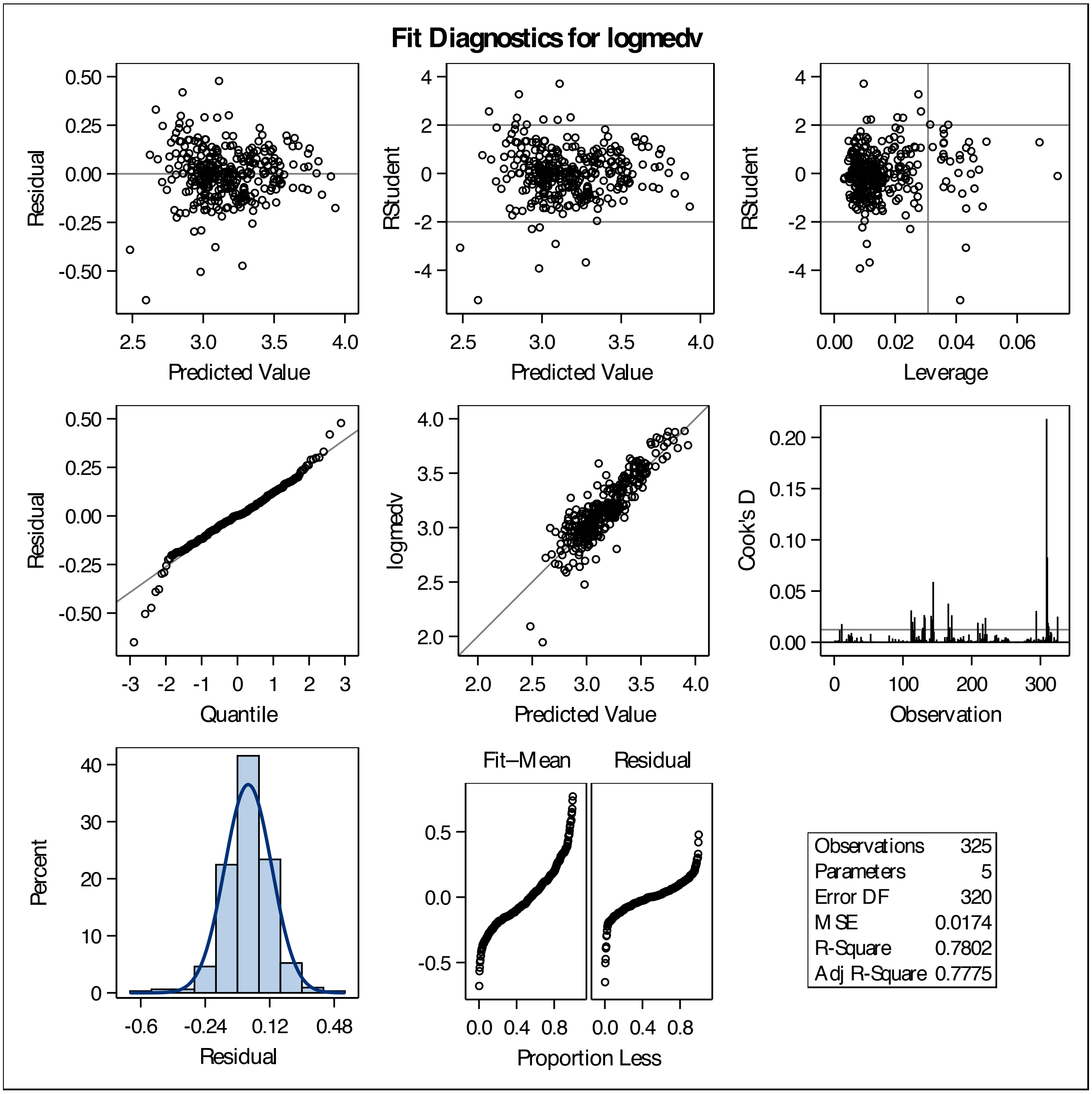
The REG Procedure
Model: MODEL1
Dependent Variable: logmedv

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|
Intercept	1	0.87577	0.11130	7.87	<.0001
age	1	-0.00238	0.00038830	-6.13	<.0001
indus	1	-0.00896	0.00171	-5.24	<.0001
nox	1	0.46342	0.17909	2.59	0.0101
rm	1	0.35470	0.01349	26.29	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: logmedv



By looking at the cook's difference, we can find that there exists some influential point in this data set. According to the context, we should exclude points whose Cook’s distance greater than 4 times the cutoff line in the plot.

After removing the influential points, I do the stepwise selection to select the best model:

The REG Procedure
Model: MODEL1
Dependent Variable: logmedv

Number of Observations Read	322
Number of Observations Used	322

Summary of Stepwise Selection

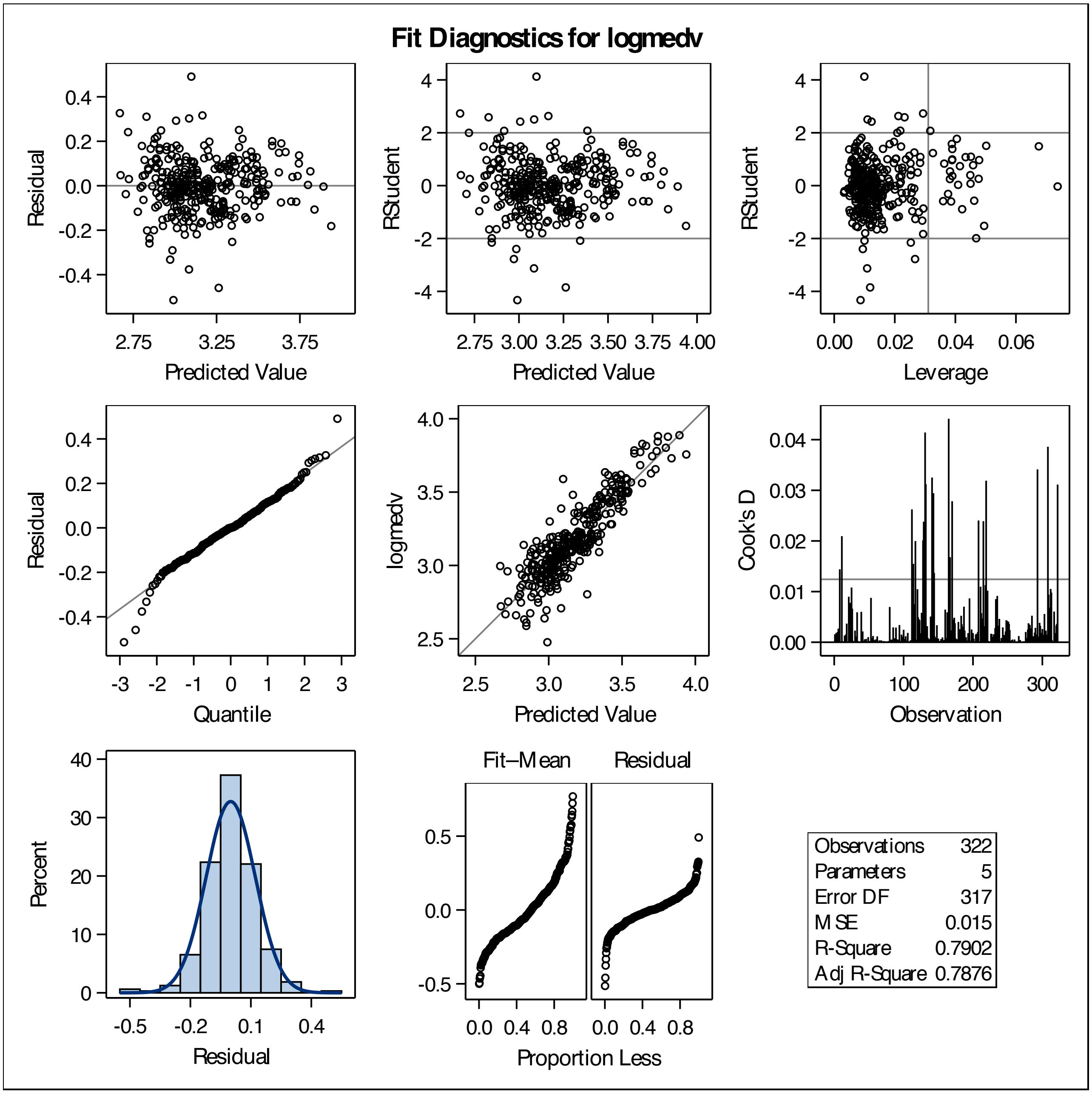
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	rm		1	0.7297	0.7297	90.5386	863.72	<.0001
2	age		2	0.0499	0.7795	17.1881	72.14	<.0001
3	indus		3	0.0058	0.7854	10.3624	8.65	0.0035
4	nox		4	0.0049	0.7902	5.0000	7.36	0.0070

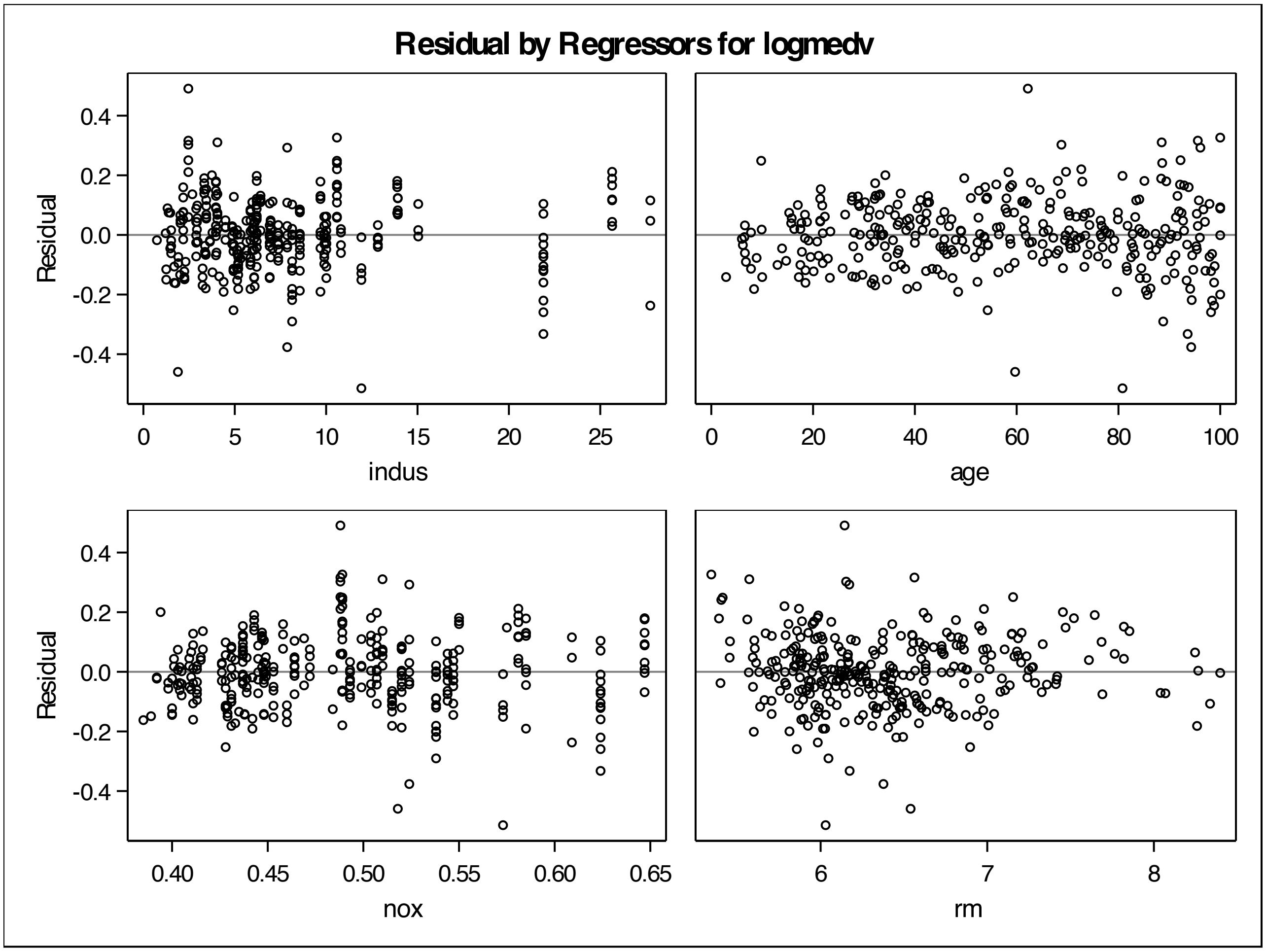
Analysis of Variance

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			

Parameter Estimates

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


Fit Diagnostics for logmedv
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
logmedv
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 322
Parameters 5
Error DF 317
MSE 0.015
R-Square 0.7902
Adj R-Square 0.7876



By looking at the cook's difference, we can tell that influential points have been excluded.

The model is significant and every predictor in this model are also significant.

According to the plot of residual, we can easily find that the residuals are normal distributed—majority of points in Rstudent graph lie between -2 and 2, histogram is symmetric, Q-Q plot is almost a straight line. So, this model is reasonable.

The model describes 79% variation in log of median home value in this subset of the data.

The relationship between the chosen predictors and median home value:

(1) rm: average number of rooms per house

The **Parameter Estimate** of rm: and the log of median home value is 0.35636. As a result, one unit of age increase will lead to the median home value become e^(0.35636) multiply the original median value. Since e^(0.35636) is larger than 1, the median value is actually increase. So, there are **positive relationship** between average number of rooms per house and the home median value

(2) Age

The **Parameter Estimate** of age and the log of median home value is -0.00247. As a result, one unit of age increase will lead to the median home value become e^(-0.00247) multiply the original median value. Since e^(-0.00247) is less than 1, the median value is actually decrease. So, there are **negative relationship** between the home age and the home median value.

(3) nox : nitric oxides concentration

The **Parameter Estimate** of nox and the log of median home value is 0.45177. As a result, one unit of age increase will lead to the median home value become e^(0.45177) multiply the original median value. Since e^(0.45177) is larger than 1, the median value is actually increase. So, there are **positive relationship** between nitric oxides concentration and the home median value.

(4) indus : proportion of non-retail business acres

The **Parameter Estimate** of age and the log of median home value is -0.0062. As a result, one unit of age increase will lead to the median home value become e^(-0.00247) multiply the original median value. Since e^(-0.0062) is less than 1, the median value is actually decrease. So, there are **negative relationship** between the home age and the home median value.