Session 19: Creating The Model

Agenda

Sl. No.	Topics For The Agenda
1.	Data Description
2.	Initial Model
3.	Residual Vs. Fitted
4.	Normal Q-Q
5.	Box Cox Transformation
6.	Residuals Vs. Fitted
7.	Normal Q-Q
8.	Comparison of "1/sqrt(PRICE)" and "log(PRICE)" Models
9.	Residuals Vs. Fitted
10.	Normal Q-Q
11.	Model2= Log(PRICE) ~
12.	Model3= Log(PRICE) ~
13.	Diagnostic Plots
14.	Model 4 After Deleting The 62nd Observation
15.	Best Subset For Model 3
16.	Predicting SAT Scores
17.	Initial Model
18.	Residual Vs. Fitted
19.	Normal Q-Q
20.	Residuals Vs. Regressors
21.	Model With Inclusion Of Square Term Of "Percent"
22.	Model With Inclusion Of Square Term Of "Expense"
23.	Residuals Vs Regressors
24.	Model With Inclusion Of Cube Term Of "Percent"
25.	Residuals Vs. Regressors
26.	Variance Decomposition Proportions

Data Description

- Data collected for several hundred used General Motors (GM) cars allows us to develop a multivariate regression model to determine car values. This is based on a variety of characteristics such as:
  - Price: suggested retail price of the used GM car
  - Mileage: number of miles the car has been driven
  - Make: manufacturer of the car such as Cadillac, Pontiac, and Chevrolet
  - Cylinder: number of cylinders in the engine
  - Liter: a more specific measure of engine size
  - Cruise: indicator variable representing whether the car has cruise control (1 = cruise)
  - Sound: indicator variable representing whether the car has upgraded speakers (1 = upgraded)
  - Leather: indicator variable representing whether the car has leather seats (1 = leather)

Initial Model

Initial Model

	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	26120	1815	14.392	< 2e-16	***	
Mileage	-0.2058	0.01857	-11.084	< 2e-16	***	
Make-Chevrolet	-17060	724.7	-23.538	< 2e-16	***	
Make-Pontiac	-18510	700.5	-26.423	< 2e-16	***	
Cylinder	-2220	501.3	-4.43	1.17E-05	***	
Liter	7691	569.3	13.509	< 2e-16	***	
Cruise1	102.4	400.7	0.256	0.798		
Sound1	227.9	387.7	0.588	0.557		
Leather1	247.2	419.8	0.589	0.556		

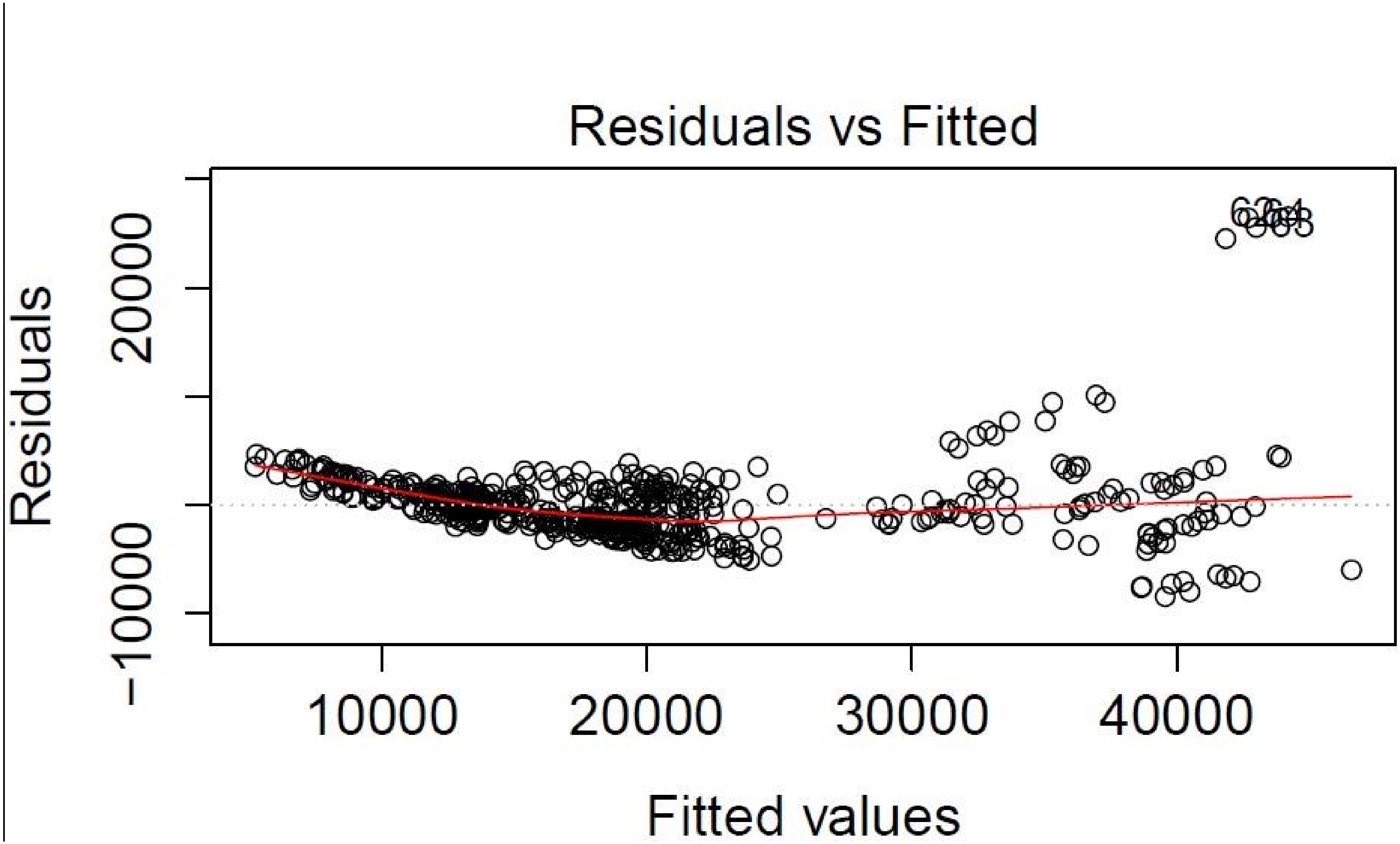
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3430 on 491 degrees of freedom

Multiple R-squared: 0.8823,   Adjusted R-squared: 0.8803

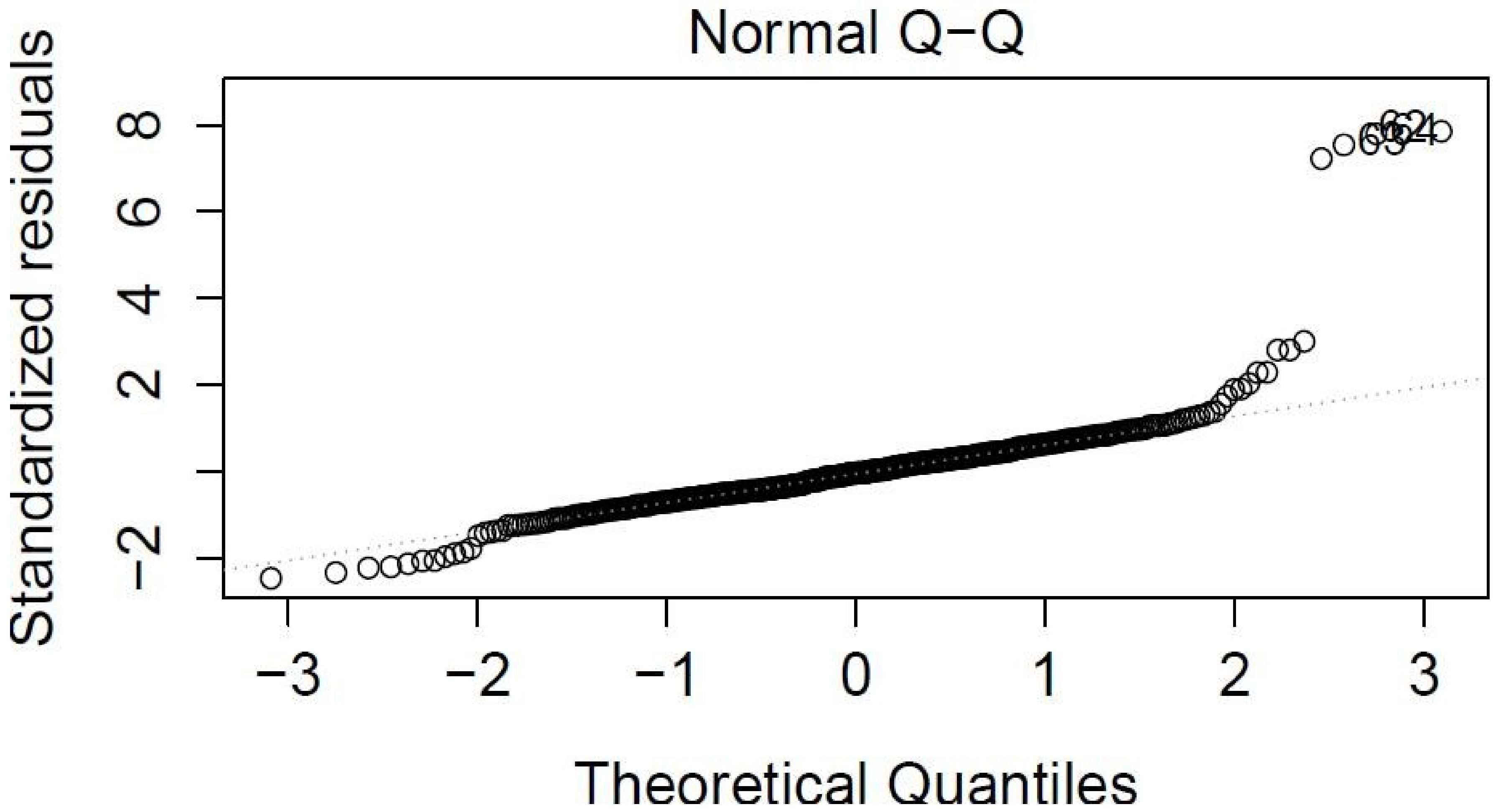
F-statistic: 459.9 on 8 and 491 DF, p-value: < 2.2e-16

Residual Vs. Fitted

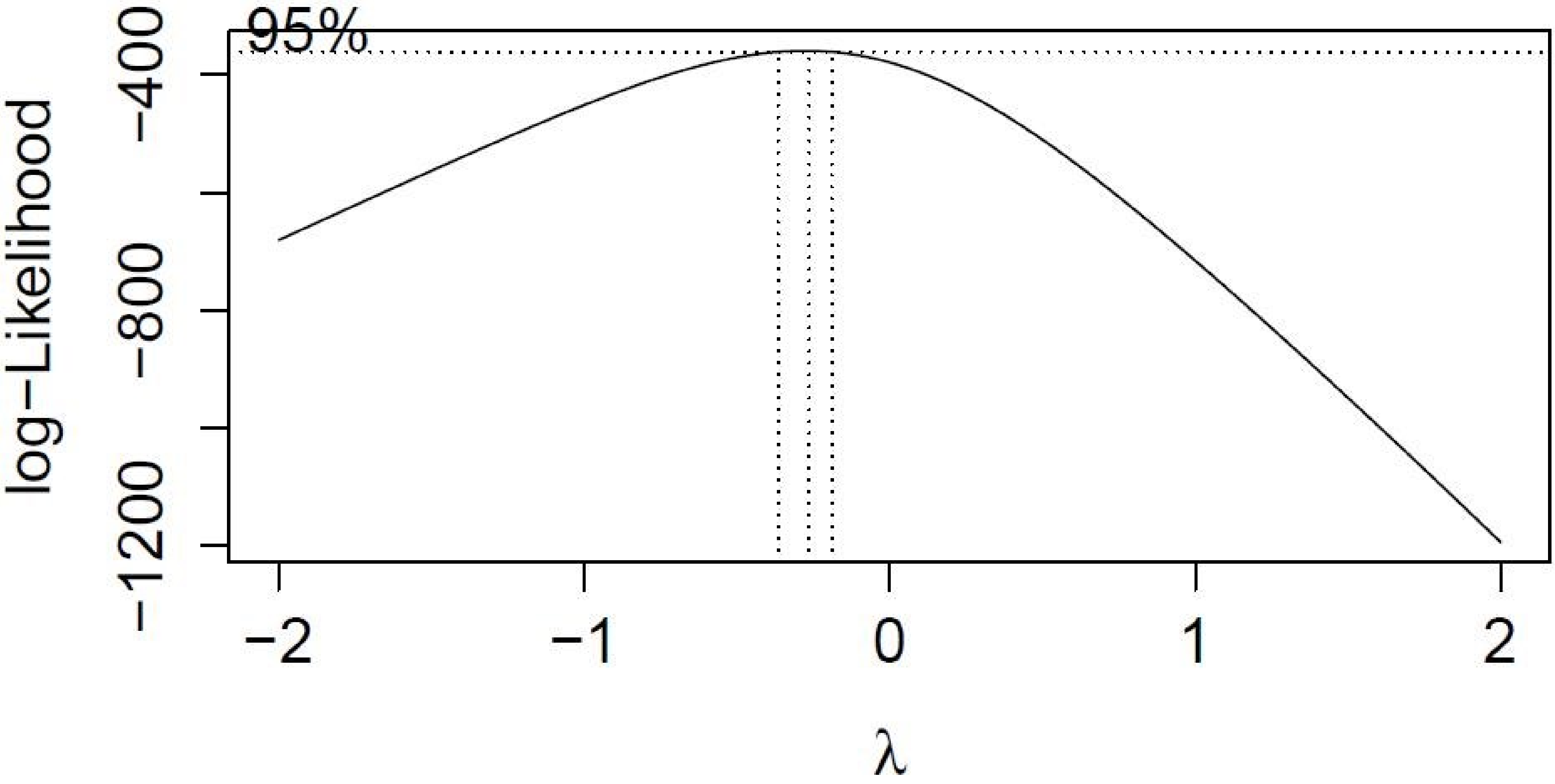


[redacted]

Normal Q-Q


Normal Q−Q
Standardized residuals
−2
2
4
6
8
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Box Cox Transformation



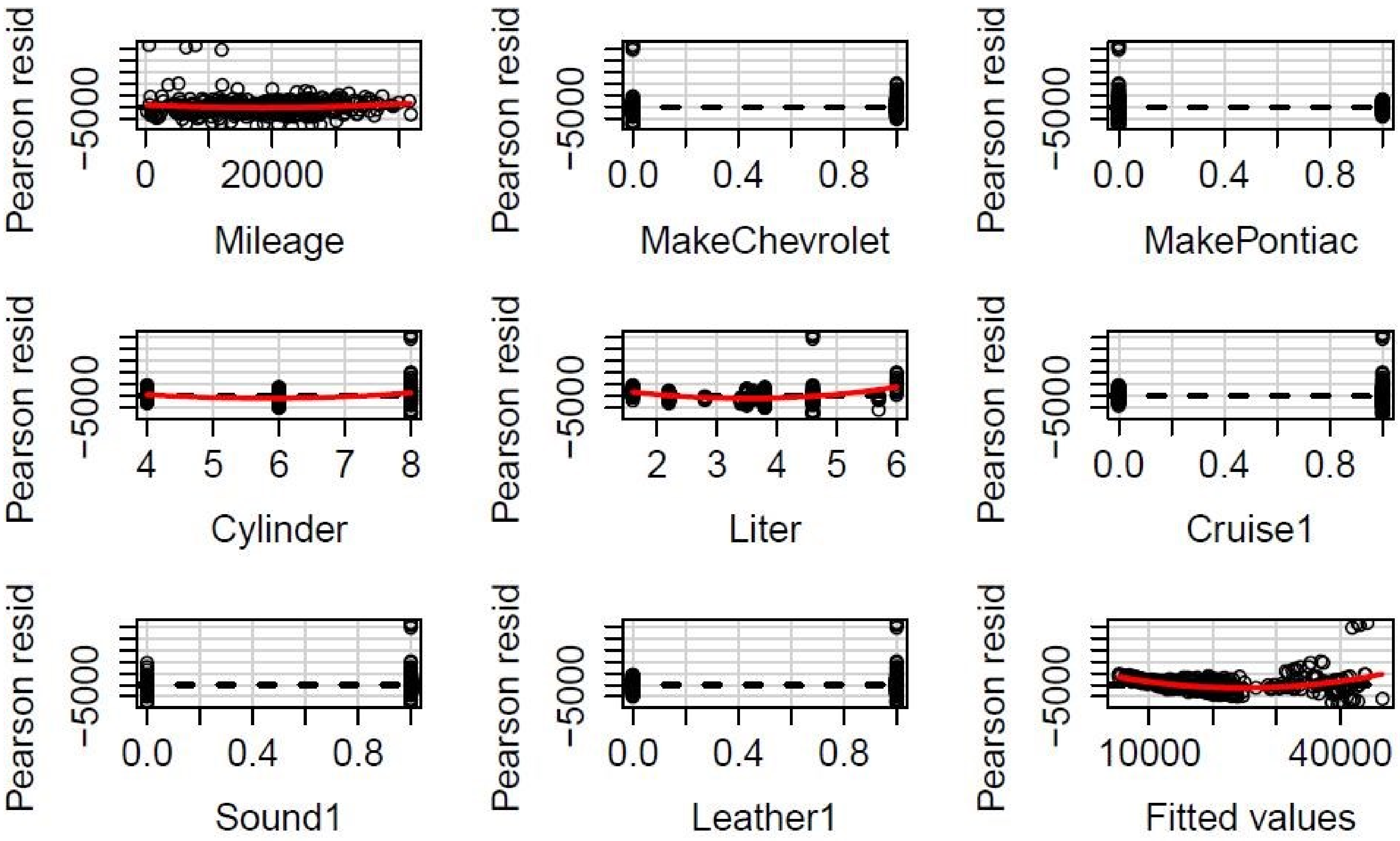
Residuals Vs. Regressors

Model with Price~

Residuals Vs Regressors

	Test stat	Pr(>|t|)
Mileage	2.012	0.045
MakeChevrolet	-0.456	0.649
MakePontiac	-0.352	0.725
Cylinder	7.235	0
Liter	9.531	0
Cruise1	-0.096	0.924
Sound1	-0.075	0.941
Leather1	-0.322	0.748
Tukey test	12.128	0

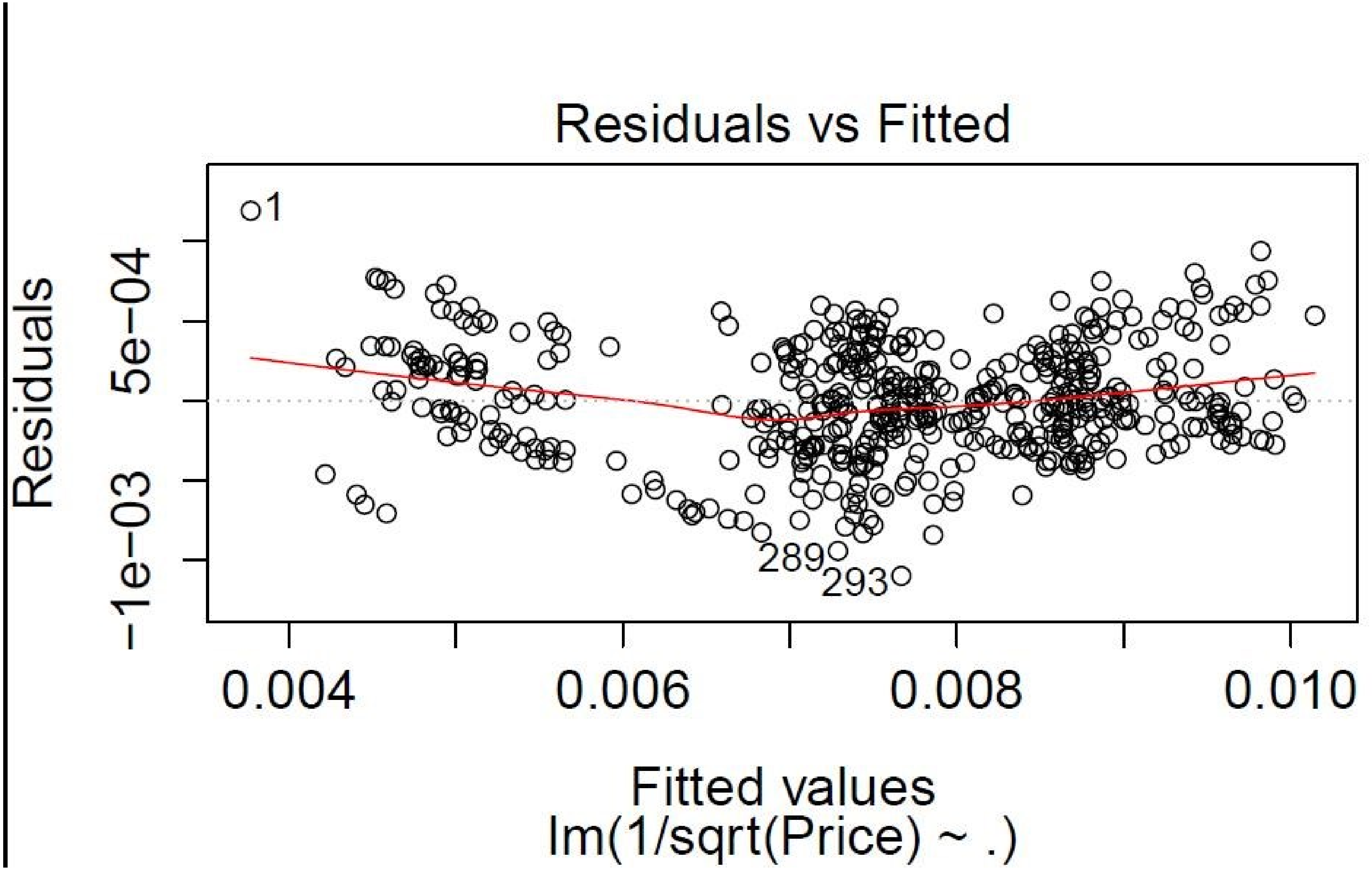
Residuals Vs. Regressors (Contd.)



Residuals Vs. Fitted

Model with 1/sqrt(Price)

There is a still a pattern

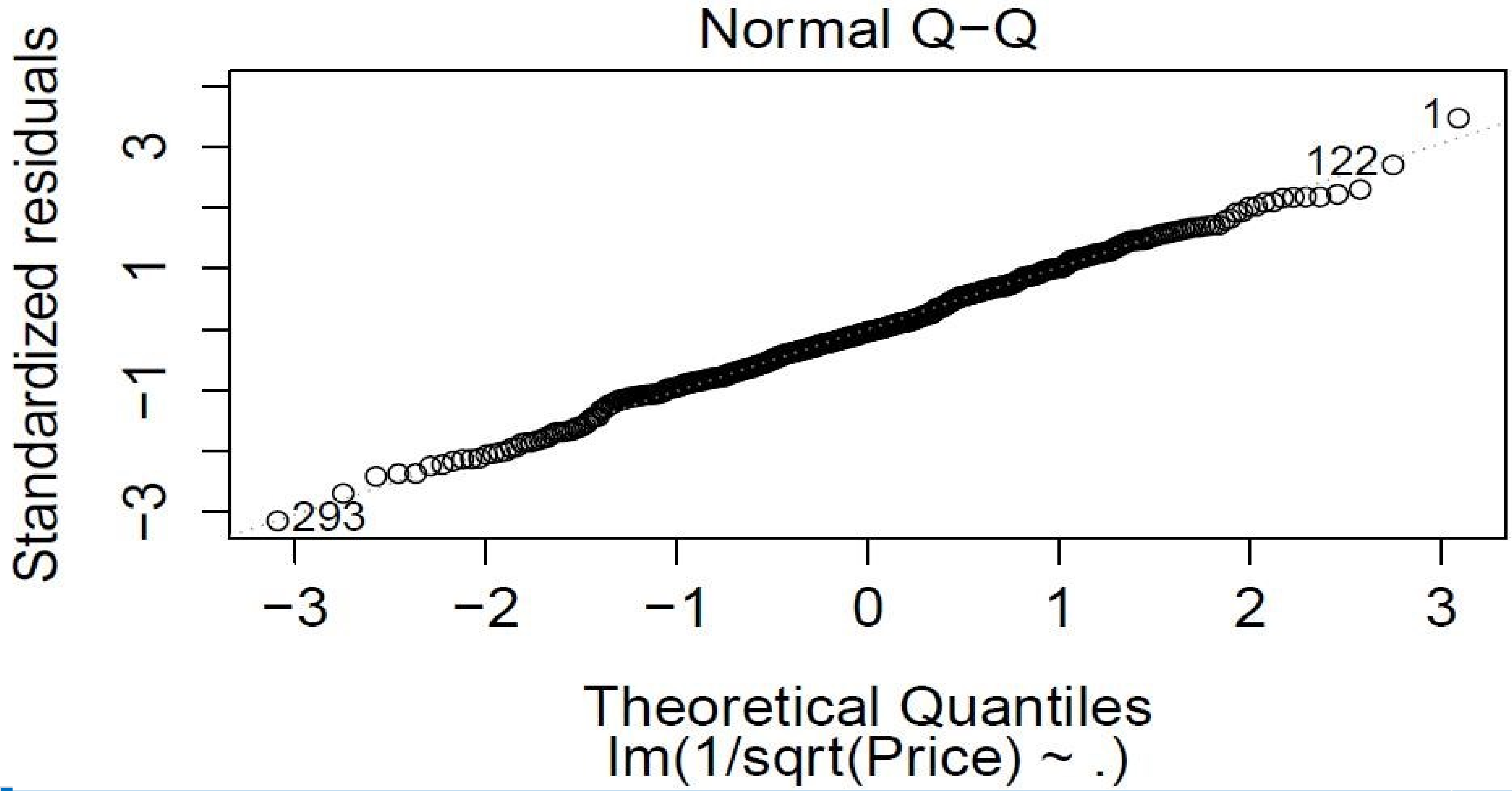


[redacted]

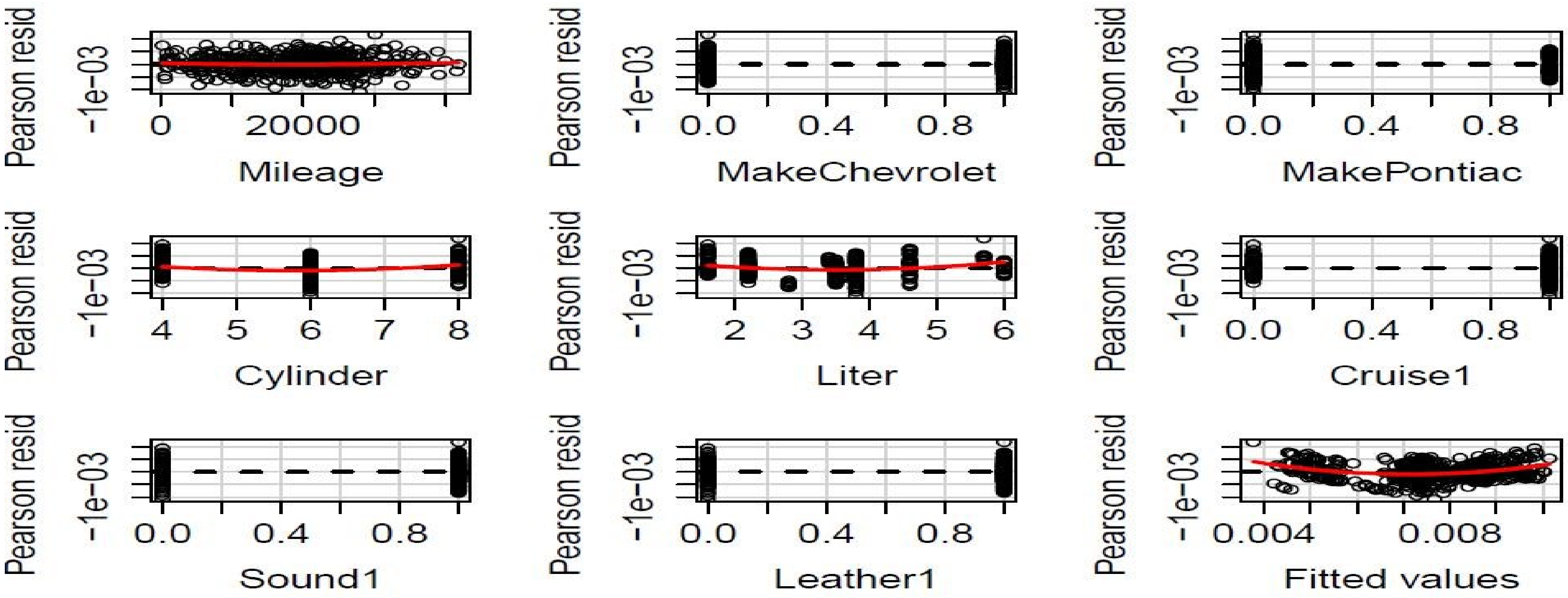
Normal Q-Q

Normality assumption is not violated

Model with 1/sqrt(Price).,



Comparison of “1/sqrt(PRICE)” and “log(PRICE)” Models



	Test stat	Pr(>|t|)
Mileage	0.863	0.389
MakeChevrolet	-1.419	0.156
MakePontiac	-1.513	0.131
Cylinder	6.598	0
Liter	5.475	0
Cruise1	1.873	0.062
Sound1	-1.737	0.083
Leather1	-2.025	0.043
Tukey test	8.27	0

Variables “Cylinder” and “Liter” are still not linear in nature

Model with 1/sqrt(Price) ~.,

[redacted]

Comparison of “1/sqrt(PRICE)” and “log(PRICE)” Models (Contd.)

Model with “1/sqrt(PRICE)”:

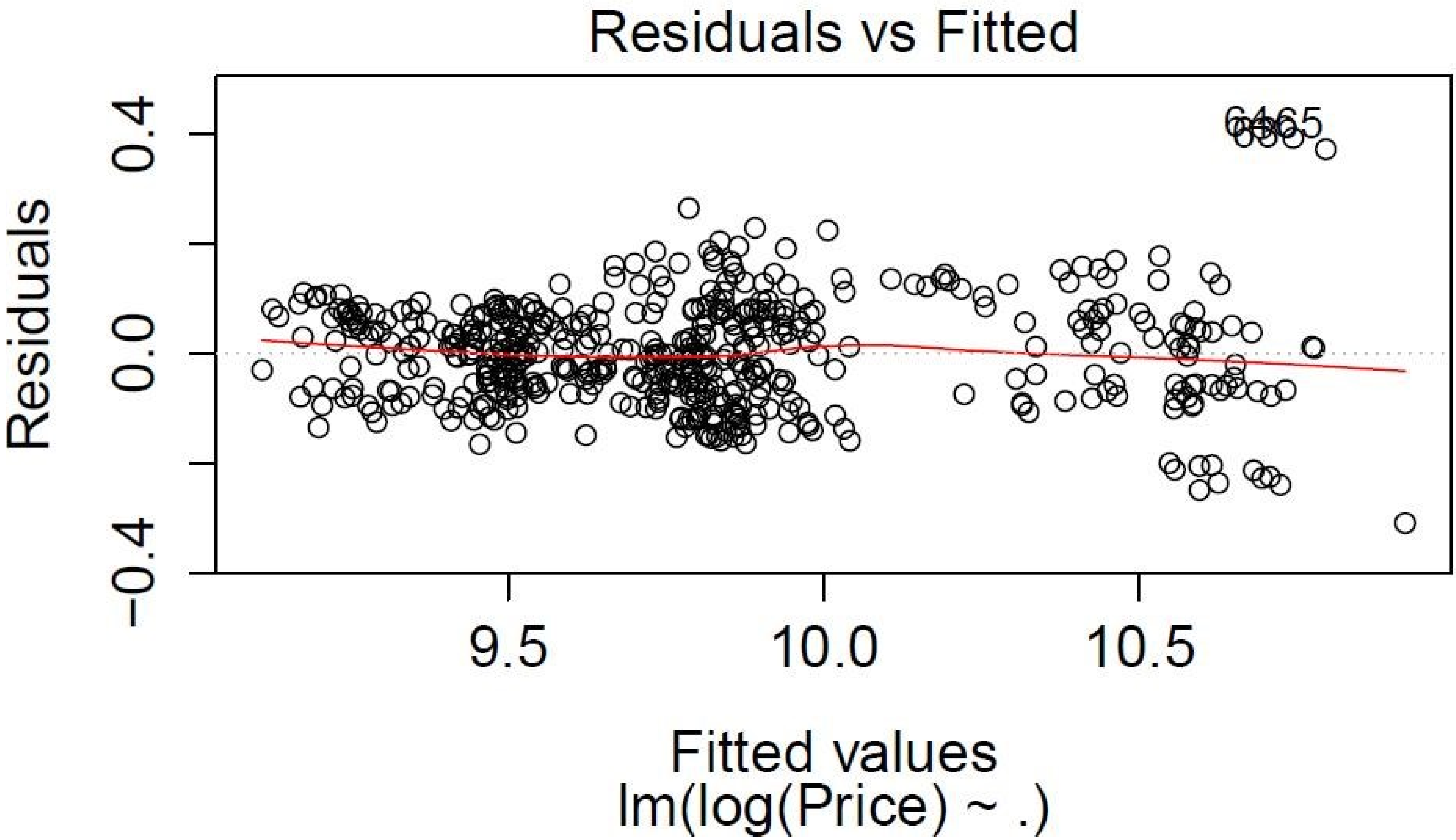
1. Residuals Vs fitted Plot is not random (Shows some curvature)
2. Normal QQ Plot is good (Assumption is not violated)
3. Residuals Vs Regressors Plot still shows a square transformation for the variables “Cylinder” and “Liter” compared to initial model

Model with “log(PRICE)”:

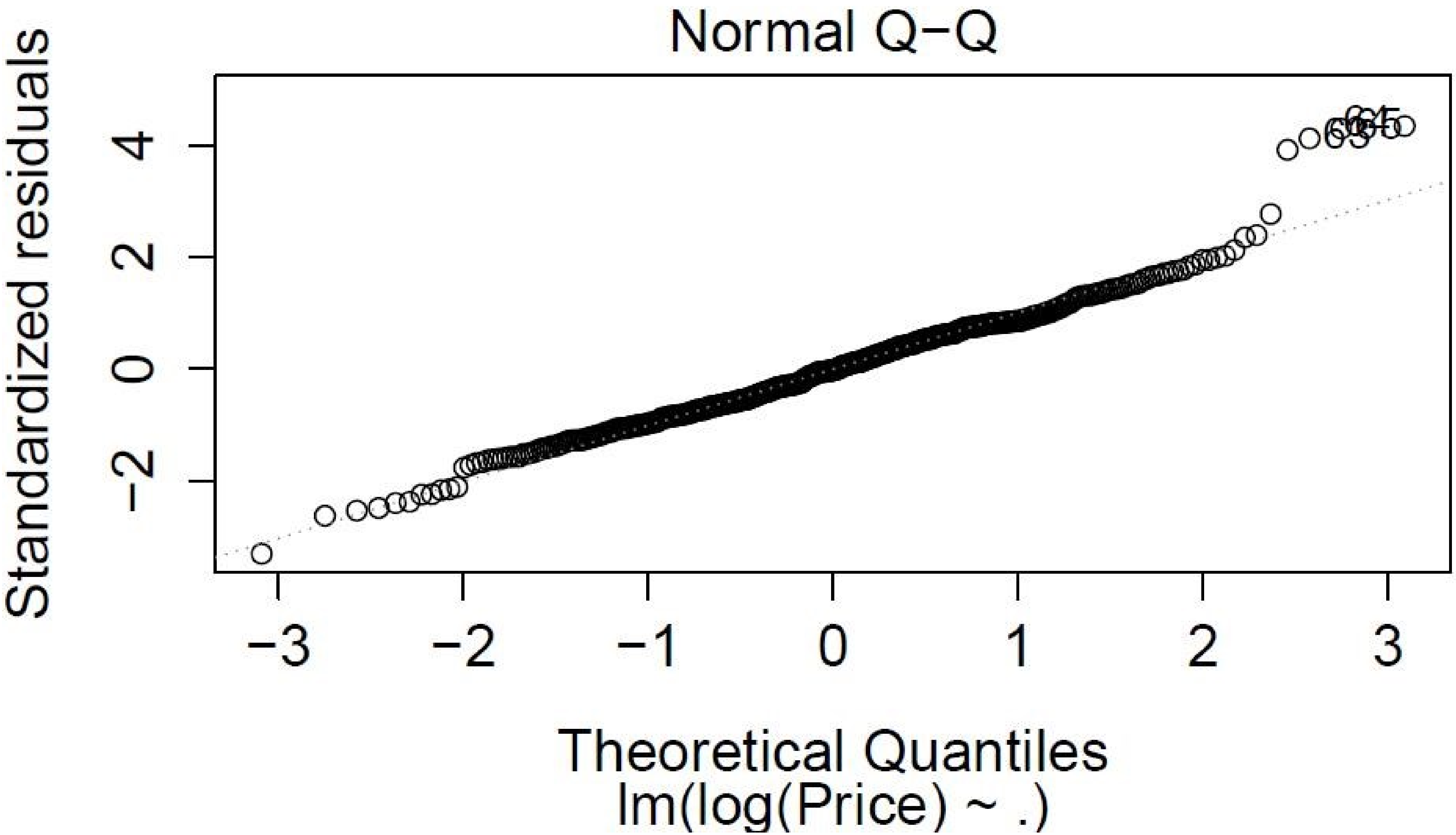
1. Residuals Vs fitted Plot is better than the initial model (Shows some randomness)
2. Normal QQ Plot is OK (Better than the initial Model)
3. Residuals Vs Regressors Plot does not show any indication of transformation for the variables “Cylinder” and “Liter” compared to initial model.

From the above observations , we can opt for log(PRICE) Model.

Residuals Vs. Fitted



Normal Q-Q


Normal Q−Q
Standardized residuals
−2
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log(Price) ~ .)

Model2= Log(PRICE) ~

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	9.862	0.05092	193.666	< 2e-16	***	
Mileage	-8.9E-06	5.21E-07	-17.045	< 2e-16	***	
MakeChevrolet	-0.6346	0.02033	-31.205	< 2e-16	***	
MakePontiac	-0.6422	0.01966	-32.671	< 2e-16	***	
Cylinder	-0.09199	0.01407	-6.54	1.55E-10	***	
Liter	0.3525	0.01598	22.062	< 2e-16	***	
Cruise1	0.01933	0.01124	1.719	0.0863	.	
Sound1	0.01999	0.01088	1.838	0.0667	.	
Leather1	0.01436	0.01178	1.219	0.2235		

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

Residual standard error: 0.09626 on 491 degrees of freedom

Multiple R-squared: 0.9471,   Adjusted R-squared: 0.9462

F-statistic: 1098 on 8 and 491 DF, p-value: < 2.2e-16

Model2= Log(PRICE) ~(Contd.)

Residuals Vs Regressors

	Test stat	Pr(>|t|)
Mileage	0.322	0.748
MakeChevrolet	0.891	0.373
MakePontiac	1.014	0.311
Cylinder	-0.673	0.501
Liter	1.468	0.143
Cruise1	-1.456	0.146
Sound1	1.252	0.211
Leather1	1.271	0.204
Tukey test	0.062	0.95

This shows that the Regressors don’t need any transformation as the p-values of all Regressors are >0.05.

Model2= Log(PRICE) ~ (Contd.)

		Variance Decompostion Proprotions								
	Condition Index	intercept	Mileage	MakeChevrolet	MakePontiac	Cylinder	Liter	Cruise1	Sound1	Leather 1
1	1	0	0.003	0.001	0.001	0	0	0.003	0.003	0.003
2	2.554	0	0	0.019	0.124	0	0	0.006	0.005	0.002
3	4.473	0	0	0.003	0.1	0	0.001	0.214	0.192	0.013
4	5.709	0	0.123	0.074	0.055	0	0.001	0.004	0.013	0.38
5	6.271	0	0.154	0	0.003	0	0	0.356	0.457	0.086
6	7.474	0.002	0.625	0.116	0.063	0.001	0.004	0.051	0.08	0.016
7	8.267	0	0.017	0.076	0.109	0.004	0.012	0.314	0.194	0.39
8	17.701	0.206	0.078	0.184	0.121	0.004	0.063	0.008	0.03	0.082
9	65.571	0.792	0	0.527	0.424	0.99	0.918	0.044	0.025	0.027

Model3= Log(PRICE) ~

Model 3 after centering the variables

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.46	0.02952	354.371	< 2e-16
Mileage	-8.9E-06	5.21E-07	-17.045	< 2e-16
MakeChevrolet	-0.6346	0.02033	-31.205	< 2e-16
MakePontiac	-0.6422	0.01966	-32.671	< 2e-16
Cruise1	0.01933	0.01124	1.719	0.0863
Sound1	0.01999	0.01088	1.838	0.0667
Leather1	0.01436	0.01178	1.219	0.2235
x	-0.09199	0.01407	-6.54	1.55E-10
y	0.3525	0.01598	22.062	< 2e-16

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

Residual standard error: 0.09626 on 491 degrees of freedom

Multiple R-squared: 0.9471,   Adjusted R-squared: 0.9462

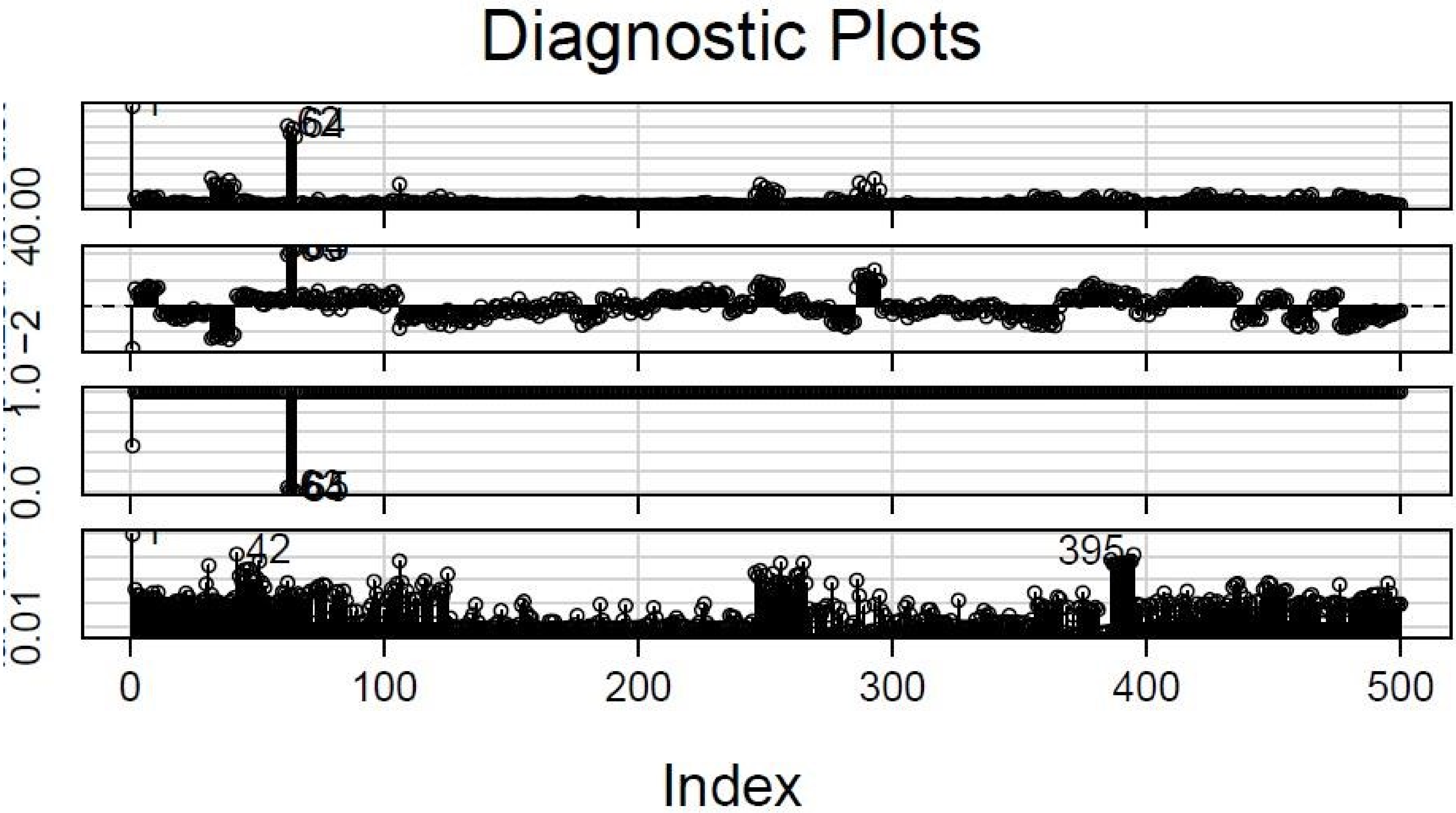
F-statistic: 1098 on 8 and 491 DF, p-value: < 2.2e-16

Model3= Log(PRICE) ~(Contd.)

		Variance Decompostion Proprotions								
	Condition Index	intercept	Mileage	MakeChevrolet	MakePontiac	Cruise1	Sound1	Leather1	x	y
1	1	0.001	0.005	0.001	0.002	0.006	0.006	0.005	0	0
2	1.632	0	0.001	0.005	0.006	0.001	0.004	0.001	0.007	0.01
3	2.543	0	0	0.01	0.183	0	0	0.001	0.003	0.005
4	4.478	0	0.037	0.004	0.005	0.331	0.198	0.127	0.001	0.003
5	5.507	0.001	0.292	0.029	0.032	0.431	0.2	0.006	0.001	0.002
6	5.61	0	0.008	0.059	0.027	0.011	0.297	0.496	0.001	0.02
7	6.689	0.004	0.464	0.176	0.13	0	0.263	0.068	0.005	0.012
8	10.034	0.127	0.177	0.002	0.005	0.079	0.003	0.175	0.222	0.237
9	21.761	0.867	0.017	0.714	0.609	0.141	0.03	0.121	0.76	0.711

Note that the collinearity has vanished.

Diagnostic Plots



Model 4 After Deleting The 62nd Observation

Coefficients:

	Estimate	Std. Error	t value	P
(Intercept)	10.45	0.02912	359.001	<
Mileage	-8.7E-06	5.16E-07	-16.838	<
MakeChevrolet	-0.6307	0.02006	-31.44	<
MakePontiac	-0.6379	0.0194	-32.878	<
Cruise1	0.01902	0.01108	1.717	0
Sound1	0.01758	0.01074	1.638	0
Leather1	0.01493	0.01161	1.286	0
x	-0.09389	0.01387	-6.77	3.
y	0.3541	0.01575	22.485	<

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

Residual standard error: 0.09485 on 490 degrees of freedom

Multiple R-squared: 0.9476,   Adjusted R-squared: 0.9467

F-statistic: 1108 on 8 and 490 DF, p-value: < 2.2e-16

Best Subset For Model 3

Start: AIC=-2331.76

log(Price) ~ Mileage + MakeChevrolet + MakePontiac + Cruise1 + Sound1 + Leather1 + x + y

	Df	**Sum of Sq**
Leather1	1	0.0138
<none>		
Cruise1	1	0.0274
Sound1	1	0.0313
X	1	0.3963
Mileage	1	2.6922
Y	1	4.5102
MakeChevrolet	1	9.0233
MakePontiac	1	9.891

Step: AIC=-2332.25

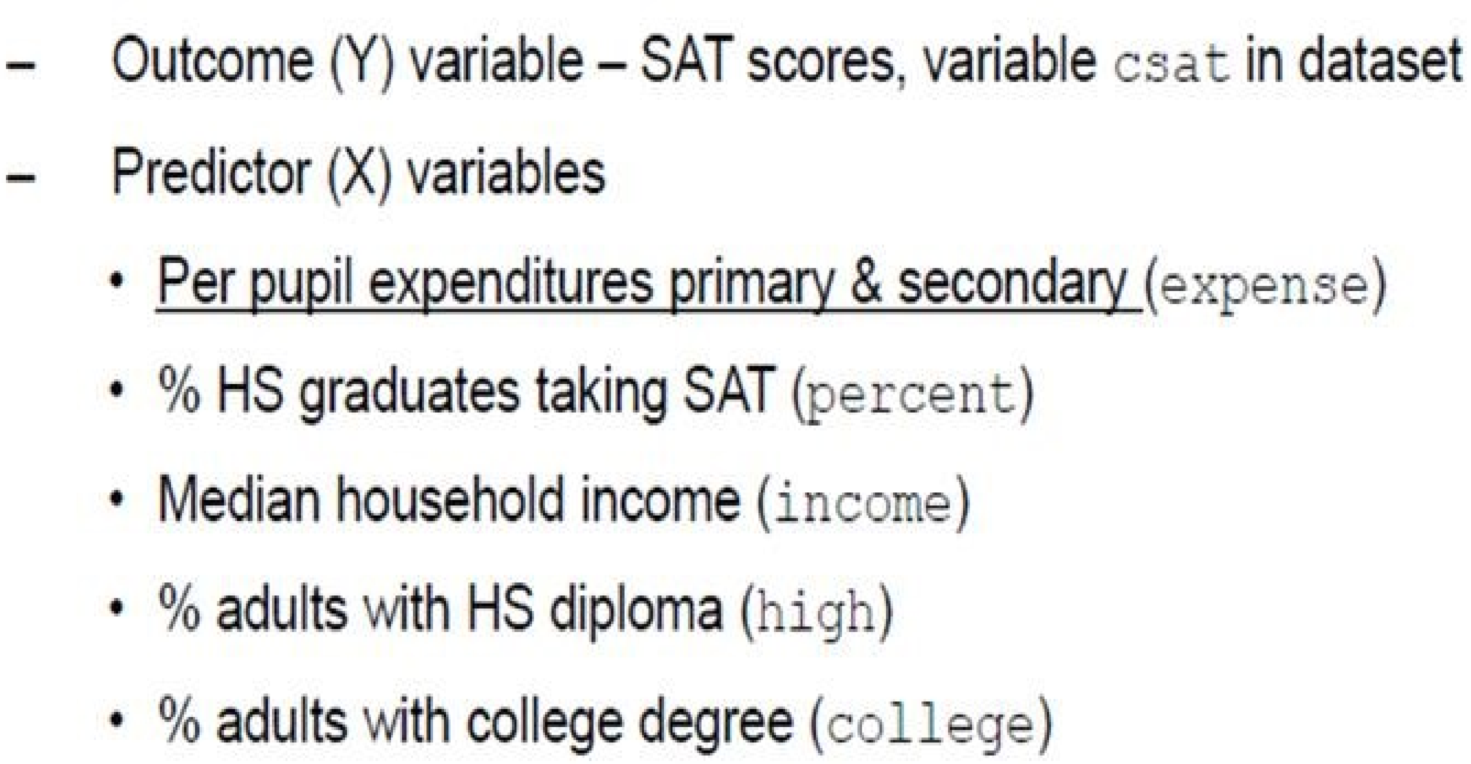
log(Price) ~ Mileage + MakeChevrolet + MakePontiac + Cruise1 + Sound1 + x + y

	Df	**Sum of Sq**
<none>		
Cruise1	**1**	**0.0217**
Sound1	1	0.0443
x	1	0.4189
Mileage	1	2.6929
y	1	4.6827
MakeChevrolet	1	9.6642
MakePontiac	1	10.9366

Predicting SAT Scores

Problem 2

- This study predicts SAT scores for 50 observations using the following factors:



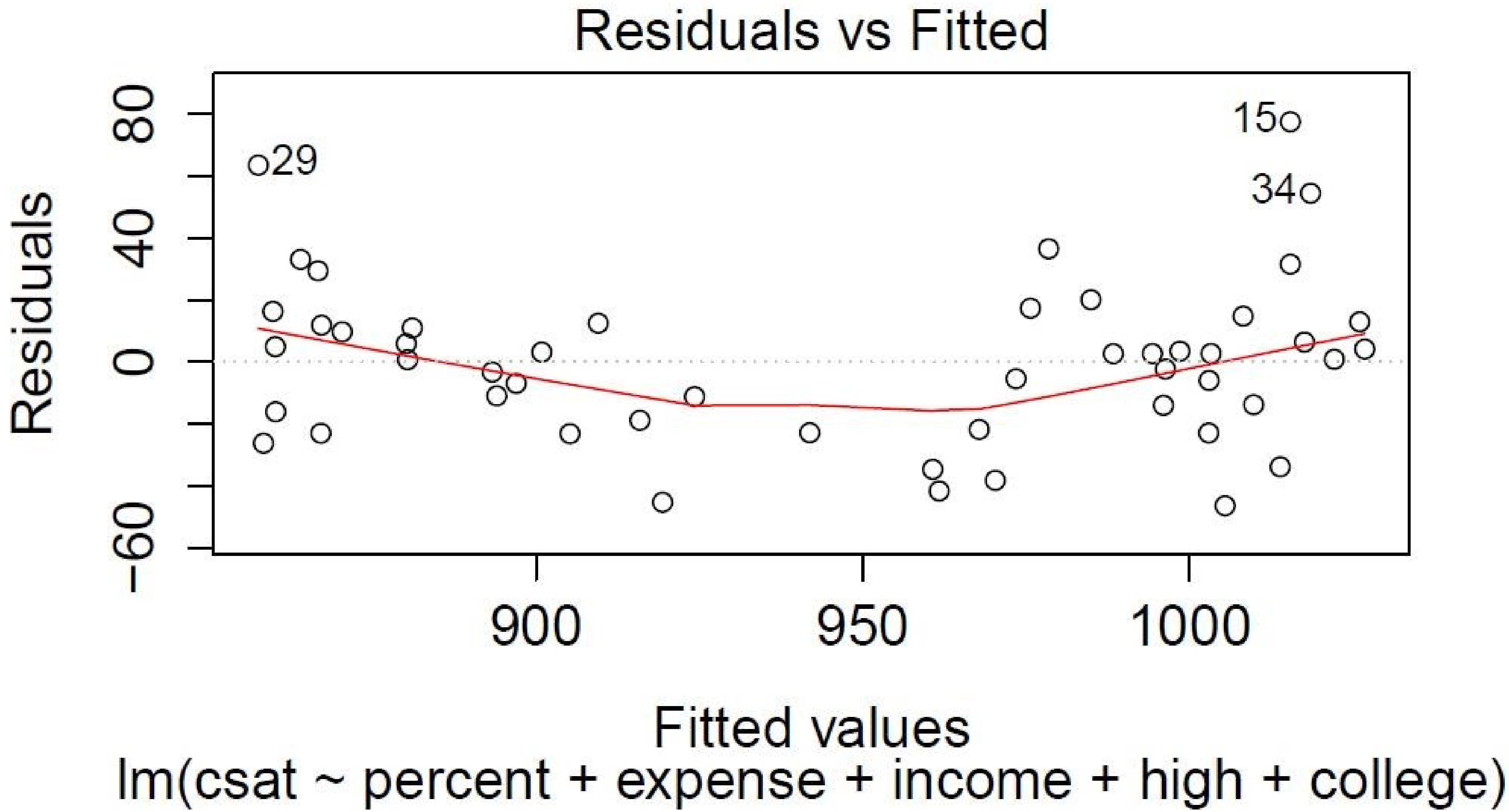
- Outcome (Y) variable – SAT scores, variable `csat` in dataset
- Predictor (X) variables
  - Per pupil expenditures primary & secondary (`expense`)
  - % HS graduates taking SAT (`percent`)
  - Median household income (`income`)
  - % adults with HS diploma (`high`)
  - % adults with college degree (`college`)

Initial Model

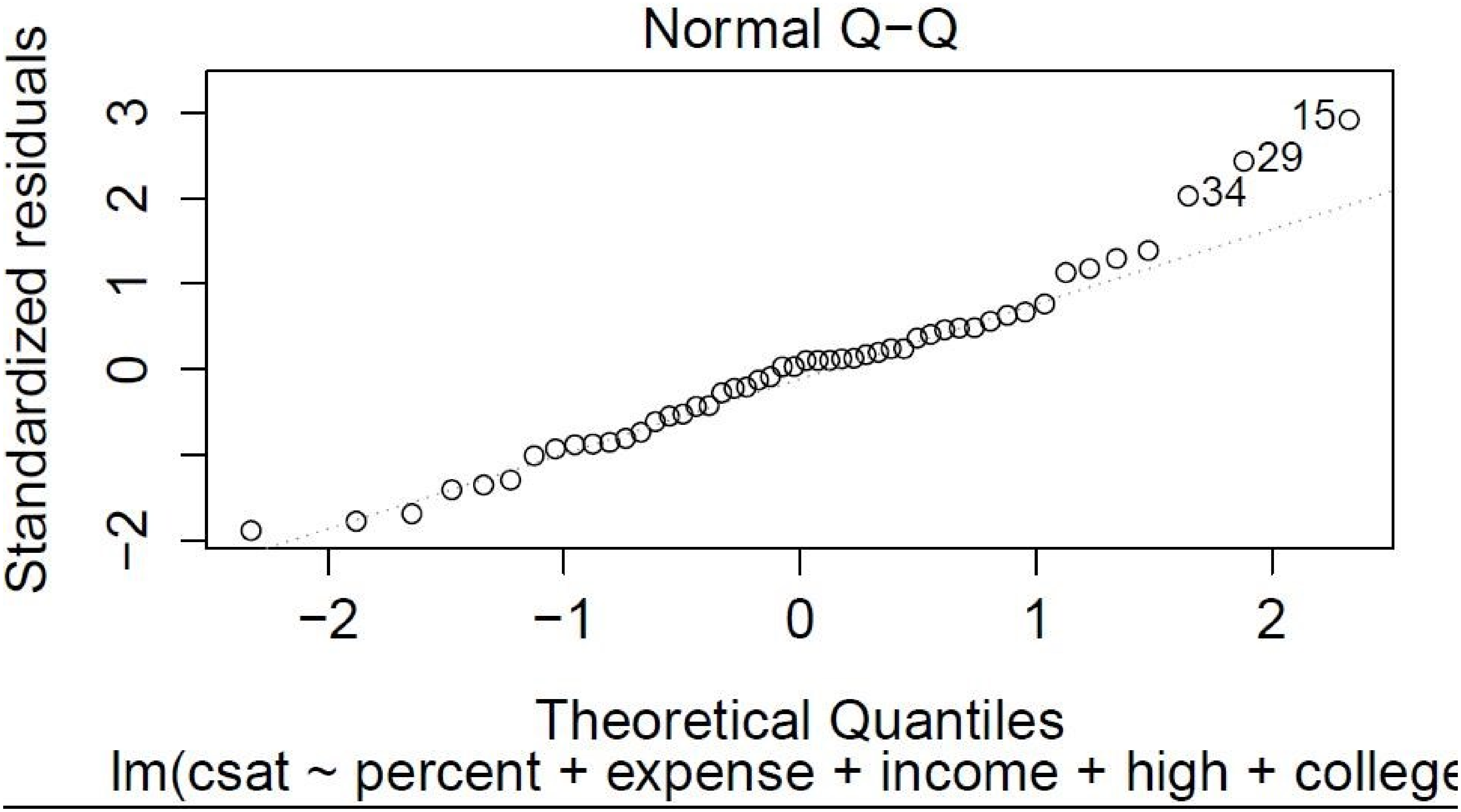
Coefficients:							
	Estimate	Std. Error	t value	Pr(>|t|)			
(Intercept)	894.4627	57.78349	15.48	< 2e-16	***		
percent	-2.76098	0.243652	-11.332	1.23E-14	***		
expense	0.009385	0.004749	1.976	0.05441	.		
income	-1.50168	1.244565	-1.207	0.23404			
high	0.510449	1.018275	0.501	0.61867			
college	5.674604	2.060506	2.754	0.00853	**		
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1							
Residual standard error: 27.7 on 44 degrees of freedom							
Multiple R-squared: 0.8414, Adjusted R-squared: 0.8234							
F-statistic: 46.69 on 5 and 44 DF, p-value: < 2.2e-16							

[redacted]

Residual Vs. Fitted



Normal Q-Q

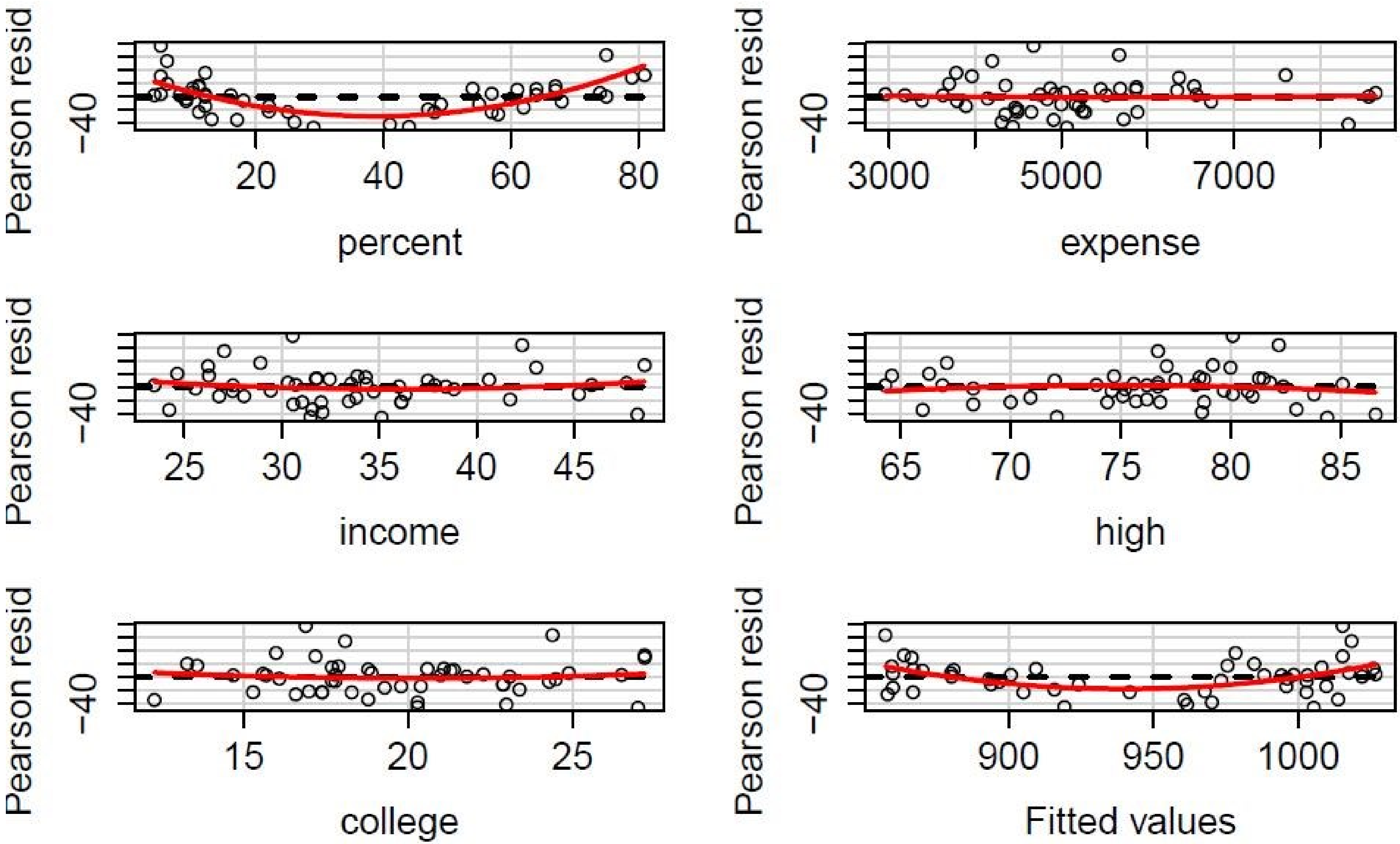

Normal Q−Q
Standardized residuals
−2
0
1
2
3
15
29
34
−2
−1
0
1
2
Theoretical Quantiles
lm(csat ~ percent + expense + income + high + college

Residuals Vs. Regressors

Residuals Vs Regressors

	Test stat	Pr(>|t|)
percent	7.547	0
expense	0.147	0.884
income	1.018	0.314
high	-0.728	0.471
college	0.636	0.528
Tukey test	3.216	0.001

Model With Inclusion Of Square Term Of “Percent”



Model With Inclusion Of Square Term Of “Percent”

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	876.457	38.412	22.817	< 2e-16	***
percent	-6.406	0.509	-12.578	5.29E-16	***
percent2	0.051	0.007	7.547	2.10E-09	***
expense	0.003	0.003	0.830	0.411	
income	-0.709	0.832	-0.852	0.399	
high	2.052	0.706	2.908	0.006	**
college	2.642	1.425	1.854	0.071	.

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

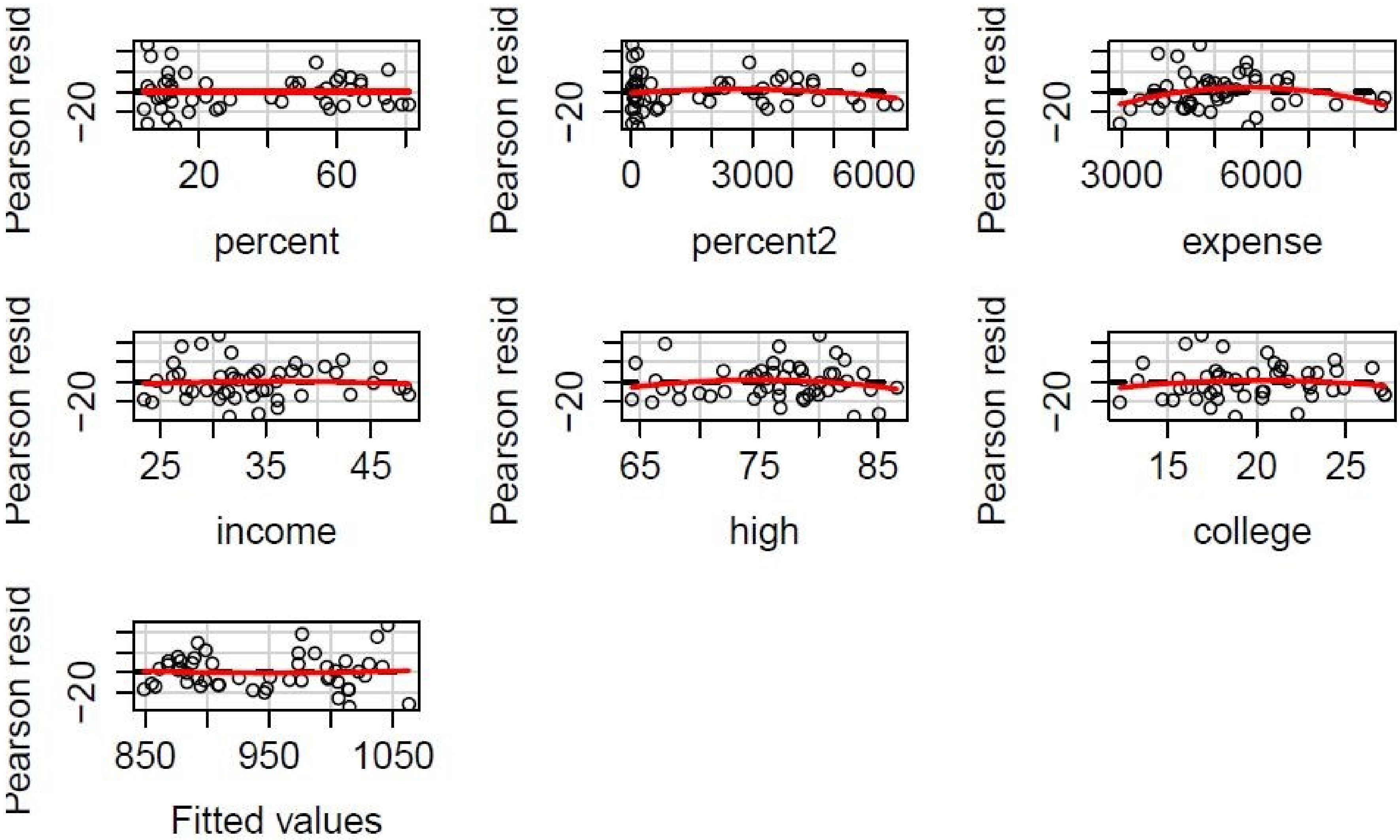
Residual standard error: 18.38 on 43 degrees of freedom

Multiple R-squared: 0.9318,   Adjusted R-squared: 0.9223

F-statistic: 97.88 on 6 and 43 DF, p-value: < 2.2e-16

[redacted]

Model With Inclusion Of Square Term Of “Expense” (Contd.)



Model With Inclusion Of Square Term Of “Expense”

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	812.9	47.19	17.228	< 2e-16	***
percent	-6.809	0.5234	-13.008	2.53E-16	***
percent2	0.05477	0.006722	8.149	3.47E-10	***
expense	0.03297	0.01437	2.295	0.0268	*
expense2	-2.6E-06	1.21E-06	-2.158	0.0367	*
income	-0.525	0.8037	-0.653	0.5172	
high	1.746	0.6923	2.523	0.0155	*
college	2.862	1.372	2.086	0.0431	*

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

Residual standard error: 17.65 on 42 degrees of freedom

Multiple R-squared: 0.9386,   Adjusted R-squared: 0.9284

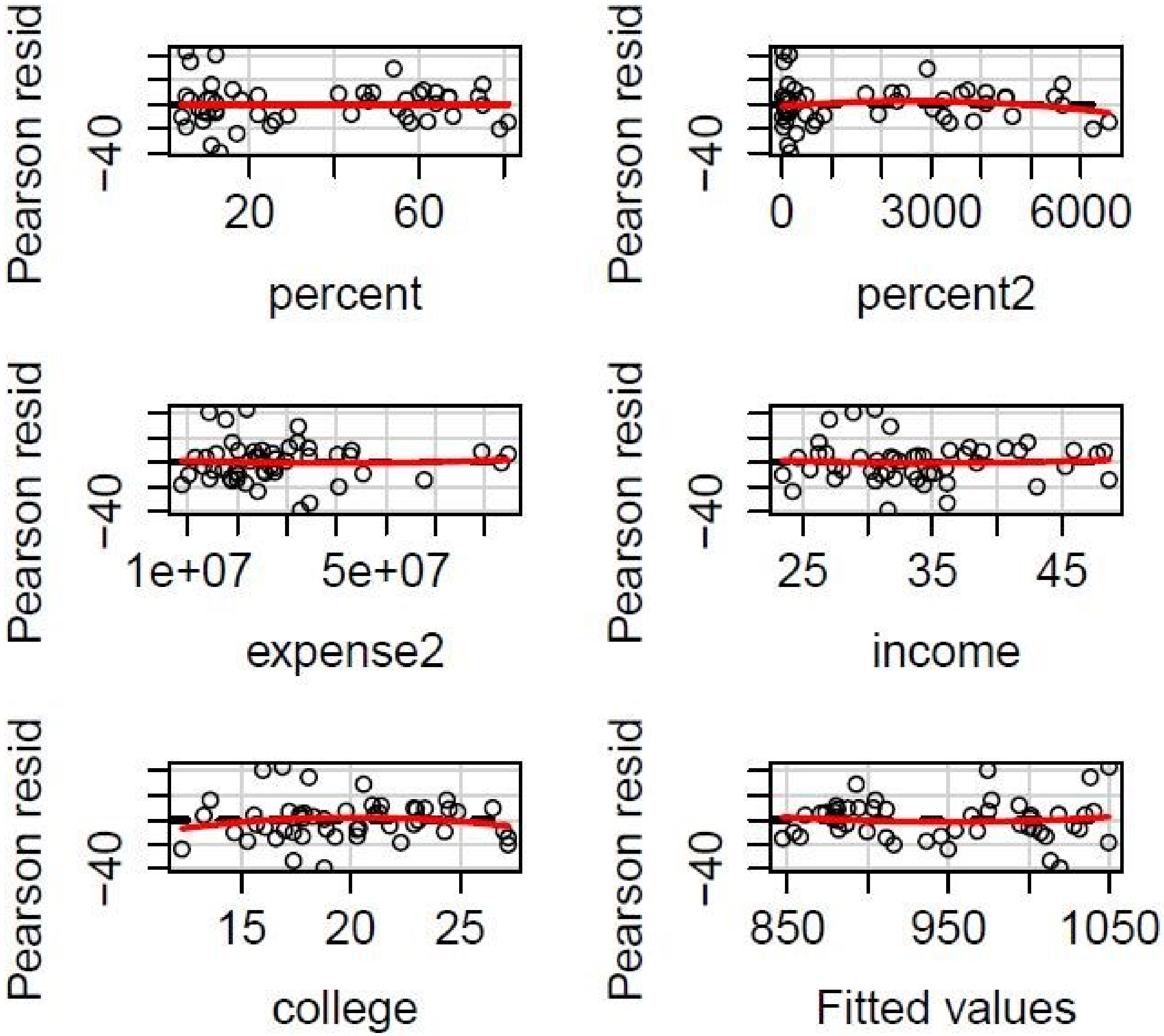
F-statistic: 91.7 on 7 and 42 DF, p-value: < 2.2e-16

Residuals Vs Regressors

Residuals Vs Regressors

	Test stat	Pr(>|t|)
percent	0.756	0.454
percent2	**-2.096**	**0.042**
expense	0.238	0.813
expense2	1.316	0.195
income	0.38	0.706
high	-0.392	0.697
college	-1.067	0.292
Tukey test	0.508	0.611

Model With Inclusion Of Cube Term Of “Percent”



Model With Inclusion Of Cube Term Of “Percent” (Contd.)

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**		
(Intercept)	814	45.28	17.976	< 2e-16	***	
percent	-9.558	1.375	-6.953	1.92E-08	***	
percent2	0.1327	0.03685	3.601	0.000847	***	
percent3	-0.00062	0.000291	-2.148	0.037661	*	
expense	0.03818	0.014	2.728	0.00934	**	
expense2	-2.9E-06	1.17E-06	-2.47	0.017748	*	
income	-0.4991	0.7713	-0.647	0.521151		
high	1.647	0.6659	2.474	0.017591	*	
college	3.319	1.333	2.489	0.016967	*	

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

Residual standard error: 16.93 on 41 degrees of freedom

Multiple R-squared: 0.9448,   Adjusted R-squared: 0.934

Residuals Vs. Regressors

	Test stat	Pr(>|t|)
percent	0.296	0.769
percent2	0.29	0.773
percent3	0.204	0.84
expense	-0.151	0.881
expense2	1.587	0.12
income	0.52	0.606
high	-0.493	0.625
college	-0.416	0.679
Tukey test	-1.148	0.251

Now , p-value of all variables are > 0.05

Variance Decomposition Proportions

	Condition Index	Variance Decomposition Proportions								
		intercept	percent	percent2	percent3	expense	expense2	income	high	college
1	1	0	0	0	0	0	0	0	0	0
2	2.934	0	0	0	0.001	0	0	0	0	0
3	8.353	0.002	0.001	0	0	0.001	0.025	0.001	0.001	0.003
4	16.865	0.001	0.03	0	0.026	0	0.001	0	0.001	0.014
5	23.418	0.061	0	0	0.001	0.005	0.001	0.096	0.006	0.217
6	36.522	0.001	0.003	0	0.002	0	0.005	0.897	0.011	0.387
7	71.916	0.482	0.001	0.001	0	0.006	0.001	0.003	0.946	0.353
8	100.494	0.439	0.006	0.024	0.036	0.858	0.873	0.001	0.029	0.006
9	155.29	0.014	0.959	0.975	0.934	0.129	0.094	0.003	0.007	0.021

```
sat$x=sat$percent-35
sat$x2=sat$percent2-1890
sat$x3=sat$percent3-117644
sat$e=sat$expense-5156
sat$e2=sat$expense2-28212245
sat$h=sat$high-76
sat$i=sat$income-34
sat$c=sat$college-20
```

Model After Centering The Variables

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**			
(Intercept)	947	2.444	387.544	< 2e-16	***		
x	-9.558	1.375	-6.953	1.92E-08	***		
x2	0.1327	0.03685	3.601	0.000847	***		
x3	-0.00062	0.000291	-2.148	0.037661	*		
e	0.03818	0.014	2.728	0.00934	**		
e2	-2.9E-06	1.17E-06	-2.47	0.017748	*		
i	-0.4991	0.7713	-0.647	0.521151			
h	1.647	0.6659	2.474	0.017591	*		
c	3.319	1.333	2.489	0.016967	*		

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

Residual standard error: 16.93 on 41 degrees of freedom

Multiple R-squared: 0.9448, Adjusted R-squared: 0.934

F-statistic: 87.72 on 8 and 41 DF, p-value: < 2.2e-16

[redacted]

Best Subset Regression Model

Start: AIC=291

csat ~ x + x2 + x3 + e + e2 + i + h + c

	Df	**Sum of Sq**	**RSS**
-I	1	120.1	1187
<none>			1175
-x3	1	1323.1	1307
-e2	1	1749.8	1350
-h	1	1755	1351
-c	1	1776.1	1353
-e	1	2133.7	1389
-x2	1	3719	1547
-x	1	13859.9	2561

Step: AIC=289.51

csat ~ x + x2 + x3 + e + e2 + h + c

	Df	**Sum of Sq**	**RSS**
<none>			1187
-x3	1	1335.9	1321
-h	1	1692.6	1356
-c	1	1749.8	1362
-e2	1	1871.1	1374
-e	1	2144	1402
-x2	1	3781.7	1565
-x	1	14233.9	2611

Next Class – Logistic Regression

Sl. No.	Topics For The Agenda
1.	Binary Response Regression Model
2.	Questions
3.	A Business Problem
4.	Linear Regression
5.	Conditional Expectation
6.	Linear Regression As Linear Probability Model
7.	Linear Regression Output Of Proposed Model
8.	Dotplot Of Predicted Probability
9.	Problems With Linear Probability Model
10.	Scatterplot: Response Variable Vs Quantitative Predictor
11.	Justification For A Sigmoid Shape
12.	Sigmoid Shape Versus Linear Shape
13.	Alternatives To Linear Probability Model
14.	Logistic Function
15.	Logistic Curve
16.	Logistic Regression
17.	Interpretation
18.	Impact Of A Regressor On Odds Ratio Is Multiplicative
19.	Impact Of A Regressor On The Probability
20.	From Log-odds To Odds Ratio
21.	Goodness Of Fit Measures
22.	Goodness Of Fit
23.	Measures Similar To R Square
24.	Confusion Matrix
25.	Goodness Of Fit
26.	R-Codes