Exploratory Analysis

A quick exploratory analysis is performed on the car dataset by plotting scatter plots of all the possible variable pairs in Figure 1 (See Appendix). The subplots being highlighted with red boxes are the scatter plots of the four potential explanatory variables versus the response variable MPG (miles per gallon), and their correlation coefficients. Overall, all the four explanatory variable candidates (VOL: cubic feet of cab space, HP: engine horsepower, SP: top speed, WT: weight) show some level of linear relationships with respect to MPG. However, there are a few outliers with low VOL values impairing the linear relationship between VOL and MPG. The scatter pattern between HP and MPG appears a slightly curved line and a few points with high HP values lying distantly from the major group of points. SP and MPG display a fairly good linear pattern; however a few outliers with low MPG value and high SP values slightly distort the linear trend. Although a few data points with low weight seem to impact the slope of the quick regression line in the plot, WT and MPG show a pretty good linearity. In summary, due to the reasonably good linear relationship between WT and MPG, no transformation is needed to correct WT. However, transformations will be applied to HP and SP in expectancy of obtaining better linear trends. Since the major problem in the scatter plot of VOL versus MPG is outliers and transformations can hardly address this problem, no transformation is performed on this variable.

Transformation to Correct Linearity

In order to preserve the linearity between MPG versus other variables and not to complicate things, transformations are only performed on explanatory variables. After exploring a few commonly used transformation functions, natural logarithm and base 10 logarithms are selected to transform shifted HP and SP respectively. The formulas are given as following:

$$tHP = \ln(HP - 35)$$

$$tSP = log_{10}(SP)$$

Another pair wise exploratory analysis scatter plot is depicted in Figure 2 (See Appendix). As it is shown, despite that the linearity trends are not perfect and there are still some outliers influencing the linear relationships, the transformations definitely ensure a better linearity between tHP and MPG, and between tSP and MPG. Nevertheless, comparing the correlation coefficients highlighted by green boxes in Figure 1 and Figure 2, the correlation coefficients after transformations are closer to -1 as compared to before transformations.

Model Fitting

In order to obtain the best model, multiple models are fitted with all the possible different combinations of (transformed) explanatory variables. The general model formula is given as following:

$$MPG = \beta_0 + \beta_1 WT + \beta_2 tHP + \beta_3 tSP + \beta_4 VOL + \epsilon$$

The fitted coefficients, corresponding P-values and R-square/adjusted R-square values are reported in Table 1.

Table 1 Regression Coefficients, Corresponding P-values and Adjusted R-squared Values for all Models

Model	Parameter	Intercept	tHP	tSP	WT	VOL	Adj. R-square
Model 1	Coefficient	93.7118	-14.2449				0.83
	P-value	<2e-16	<2e-16				
Model 2	Coefficient	330.52		-144.9			0.5158
	P-value	2.76e-16		3.08e-14			
Model 3	Coefficient	68.16545			-1.11222		0.8192
	P-value	<2e-16			<2e-16		
Model 4	Coefficient	50.22				-0.16637	0.1359

	P-value	<2e-16				6.56e-4	
Model 5	Coefficient	-113.549	-22.537	118.243			0.8895
	P-value	4.12e-4	<2e-16	2.19e-9			
Model 6	Coefficient	84.68715	-7.94483		-0.56544		0.8763
	P-value	<2e-16	1.71e-8		2.09e-7		
Model 7	Coefficient	98.929	-13.6377			-0.07866	0.8553
	P-value	<2e-16	<2e-16			1.32e-4	
Model 8	Coefficient	134.1133		-34.4067	-0.9663		0.83
	P-value	6.37e-7		9.19e-3	<2e-16		
Model 9	Coefficient	348.71306		-145.61		-0.16943	0.648
	P-value	<2e-16		<2e-16		2.03e-7	
Model 10	Coefficient	6838761			-1.101	-0.0107	0.8151
	P-value	<2e-16			<2e-16	0.648	
Model 11	Coefficient	-66.6238	-16.6041	88.3377	-0.3443		**0.903**
	P-value	0.03988	2.56e-11	8.63e-6	8.85e-4		
Model 12	Coefficient	-87.2469	104.2547	-21.3483		-0.02687	**0.8907**
	P-value	0.0183	1.86e-6	<2e-16		0.1776	
Model 13	Coefficient	157.0339		-44.8615	-0.87642	-0.0434	0.8345
	P-value	2.13e-7		0.00196	1.22e-14	0.07854	
Model 14	Coefficient	89.20146	-8.6983		-0.46707	-0.04439	0.8827
	P-value	<2e-16	1.69e-9		2.21e-5	0.0238	
Model 15	Coefficient	-55.5715	-16.2789	82.3857	-0.3285	-0.01348	**0.9023**

	P-value	0.12282	1.27e-10	1.08e-4	1.937e-3	0.48303	

The models with top three (Model 11, Model 15, Model 12) adjusted R-squared values are highlighted with bold font. Post model-fit diagnostics are focused on these three models.

Model Diagnostics

The desired model is going to be used for inferential purpose, therefore, besides the linearity assumption, further assumptions on error ϵ need to be satisfied: $\varepsilon|x \sim Norm(0, \sigma^2)$. Therefore post-model-fit-diagnostics are performed here to check the assumption satisfactions.

Model 11 Diagnostics

In Figure 3 (See Appendix), the residual plots indicate heteroscedasticity and a violation of error assumptions. The subplot 2 and subplot 3 in Figure 3 demonstrates that the heteroscedasticity mainly come from tHP and tSP, and is partially contributed by WT. This issue is addressed by trying out different transformation functions on the response variable MPG, and then therefore trying new transformations on explanatory variables to correct distorted linearity. After trying out a few different transformation functions, a natural log transformation is performed on the response variable, and a new transformation is applied to SP as well.

$$tMPG = \ln(MPG)$$

$$tHP = \ln(HP - 35)$$

$$tSP_{new} = -\frac{1}{SP - 80}$$

A new model is constructed based on new transformed variables:

$$tMPG = \beta_0 + \beta_1 WT + \beta_2 tHP + \beta_3 tSP_{new} + \epsilon$$

The newly obtained model comes with an adjusted R-squared of 0.9289, and the formula is given by:

$$\ln(MPG) = 4.57 - 0.0268 * WT - 2.70 * \left(-\frac{1}{SP-80}\right) - 0.086 * \ln(HP - 35)$$

Figure 4 (See Appendix) is the residual plot for the newly obtained model, as it is shown, there is no apparent pattern embedded in the residuals, although there are some outliers dwell in the outskirt of the main residual cluster region. Figure 5 (See Appendix) shows outlier diagnosis and some other model diagnosis for Corrected Model 11. As it is labeled out, a few observations (such as Obs. No. 1, Obs. No. 8, Obs. No. 29, Obs. No. 51, Obs. No. 82) have large cook's distance values, and high leverage values, and they are the main resources for residuals to deviate from theoretical quintiles.

Model 15 Diagnostics

Figure 6 (See Appendix) shows the residual diagnosis plots for Model 15. It is apparent that the residuals are more spread out around zero with the increasing fitted values; this is an evidence of heteroscedasticity. The presence of heteroscedasticity is addressed by transforming response variable MPG, and then further on transforming explanatory variables to correct linearity. In the end a natural log transformation is performed on the response variable and transformations are applied to SP and VOL as well.

$$tMPG = \ln(MPG)$$

$$tHP = \ln(HP - 35)$$

$$tSP_{new} = -\frac{1}{SP-80}$$

$$tVOL = \ln(VOL)$$

A new model is constructed based on new transformed variables:

$$tMPG = \beta_0 + \beta_1 WT + \beta_2 tHP + \beta_3 tSP_{new} + \beta_4 tVOL + \epsilon$$

The newly obtained model comes with an adjusted R-squared of 0.9284, and the formula is given by:

$$\ln(MPG) = 4.72 - 0.0269 * WT - 2.43 * \left(-\frac{1}{SP-80}\right) - 0.10 * \ln(HP-35) - 0.0245 * \ln(VOL)$$

Figure 7 (See Appendix) is the residual plot for the newly obtained model. There is no apparent pattern embedded in the residuals. However there are a few outliers. Figure 8 (See Appendix) shows outlier diagnosis for Corrected Model 15. As it is labeled out, a few observations (such as Obs. No. 1, Obs. No. 8, Obs. No. 29, Obs. No. 55, Obs. No. 82) with large cook's distance values and high leverage values are identified. Some of these outliers are also responsible for driving residuals deviate from theoretical quintiles.

Model 12 Diagnostics

Figure 9 (See Appendix) shows the residual diagnosis plots for Model 12. The first subplot displays the residual vs. fitted value, it is mostly evenly and randomly spread around residual=0 line, however it is slightly more spreading out when fitted value gets larger. However in the other subplots which display scatter patterns of residuals versus explanatory variables, it appears that residuals are more compactly centered in the middle-value ranges of tHP and tSP, and more widely scatter around in low-value ranges and high value ranges of tHP and tSP. Meanwhile, the residual versus VOL provides concrete evidence of outliers. The heteroscedasticity issue is addressed by transforming response variable MPG and explanatory variables. The transformation functions are given as following.

$$tMPG = \ln(MPG)$$

$$tHP = \ln(HP-35)$$

$$tSP_{new} = -\frac{1}{SP-80}$$

$$tVOL = \ln(VOL)$$

The newly constructed model is given by:

$$tMPG = \beta_0 + \beta_2 tHP + \beta_3 tSP_{new} + \beta_4 tVOL + \epsilon$$

The newly obtained model comes with an adjusted R-squared of 0.8881, and the formula is given by:

$$\ln(MPG) = 7.00 + 10.65 * \left(-\frac{1}{SP - 80}\right) - 0.628 * \ln(HP - 35) - 0.1124 * \ln(VOL)$$

Figure 10 (See Appendix) shows residual plots for the newly obtained model. There is no apparent pattern embedded in the plot of residuals versus fitted value. Even though it is not perfect, the heteroscedasticity in the second and third subplots are much better as compared to the subplots in Figure 9 (See Appendix). However the outliers presented in residuals versus tVOL in Figure 9 (See Appendix) still reside outside of the main data point cluster. Figure 11(See Appendix) demonstrates outlier diagnosis for Corrected Model 12. Outliers are identified and labeled by observation numbers. Interestingly, in this fitted model, the identified outliers are quite different from Corrected Model 11(See Appendix) and Corrected Model 15(i.e. Obs. No. 30, Obs. No. 55, Obs. No. 72, and Obs. No. 82). These outliers obtain large cook's distances and large leverage values. Moreover they drive the two tails of residual populations deviate from theoretical quintiles.

Model Selection

The final model is selected based on the adjusted R-squared values and AIC values. Details are listed out in Table 2. Evidently, Corrected Model 11 has the highest adjusted R-squared and the lowest AIC value, meanwhile it only fits 3 explanatory variables to the model. Therefore Corrected Model 11 is selected to serve as the final model for predicting MPG and related inference purposes.

Table2 Adjusted R-Squared and AIC for the 3 Corrected Models

Models	Adjusted R-squared	AIC	No. of Explanatory Variables
Corrected Model 11	**0.9289305**	**-173.291**	**3**
Corrected Model 15	0.9284017	-171.7412	4
Corrected Model 12	0.8880915	-136.0611	3

Summary

After exploring the provided dataset, applying various transformation functions and fitting multiple regression models, a regression model using MPG as response variable and HP, SP and WT as explanatory variables is constructed for inference purposes. The final selected regression model formula is given as following:

$$\ln(MPG) = 4.57 - 0.0268 * WT - 2.70 * \left(-\frac{1}{SP - 80}\right) - 0.086 * \ln(HP - 35)$$

It is worth mention that the constructed model has an adjusted R-squared of 0.9289 and an AIC value of -173.291. The post-model-diagnosis plots (Figure 4 and Figure 5) indicate that there are a few observation points with large cook's distance values and high leverages (identified observation number of these data points are: Obs. No. 1, Obs. No. 8, Obs. No. 29, Obs. No. 55, Obs. No. 82) dwell outside of the regression trend, these observation points can potential impair the prediction quality and inference accuracy of the constructed model.

Appendix

Figure 1 Exploratory Analysis

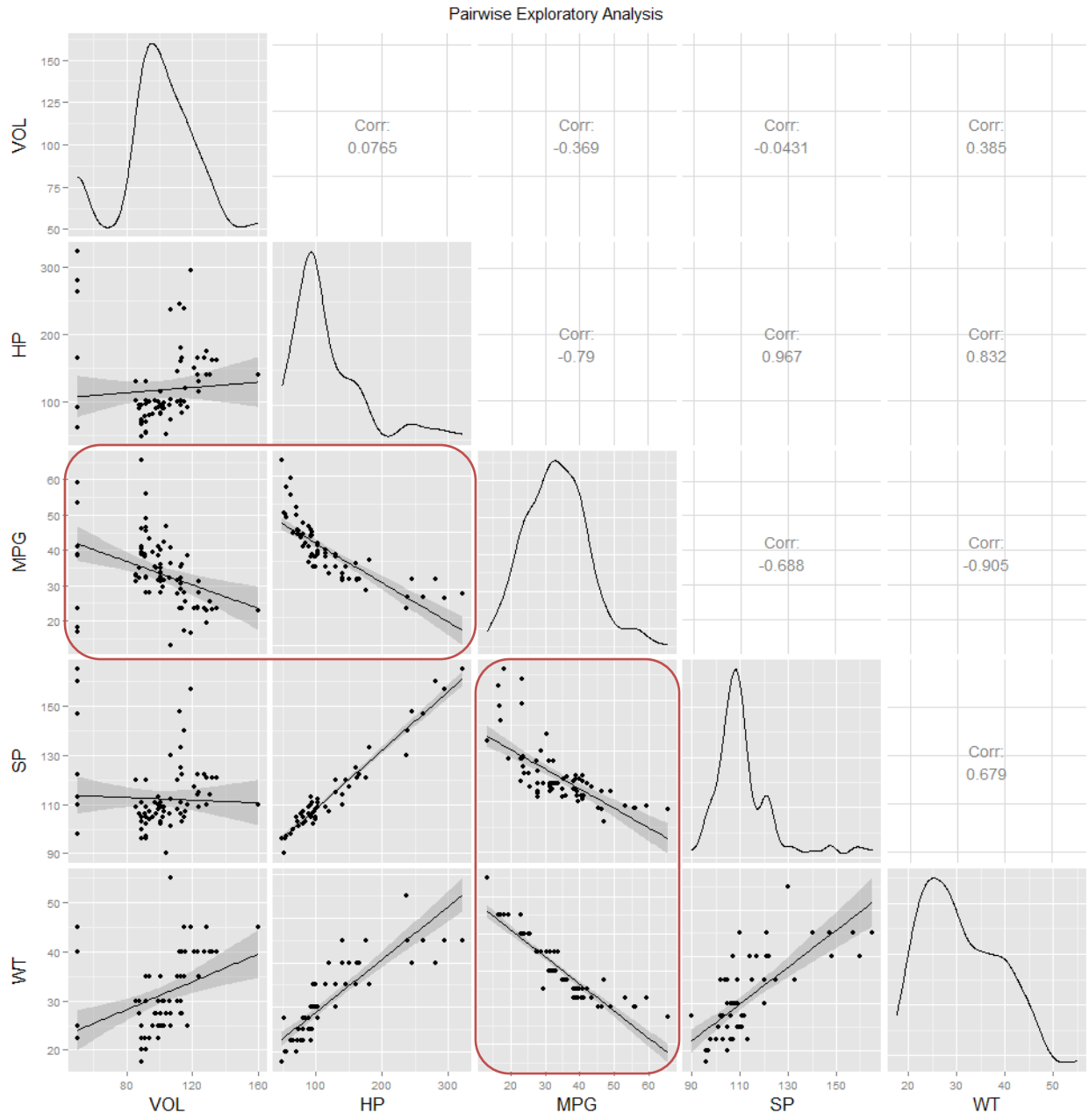

Figure 2 Exploratory Analyses after Transforming Explanatory Variables

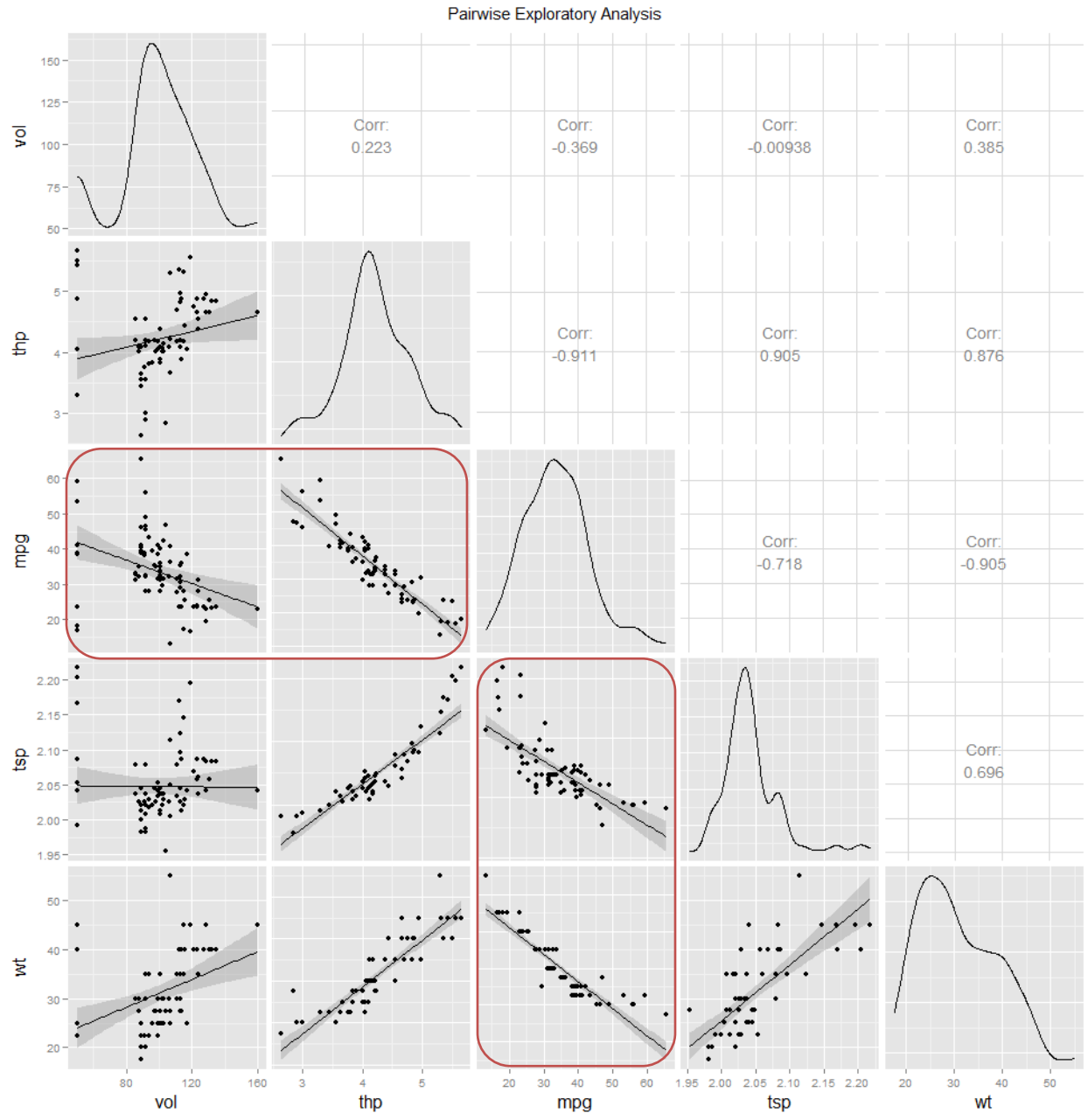

Figure 3 Residual Plots for Model 11

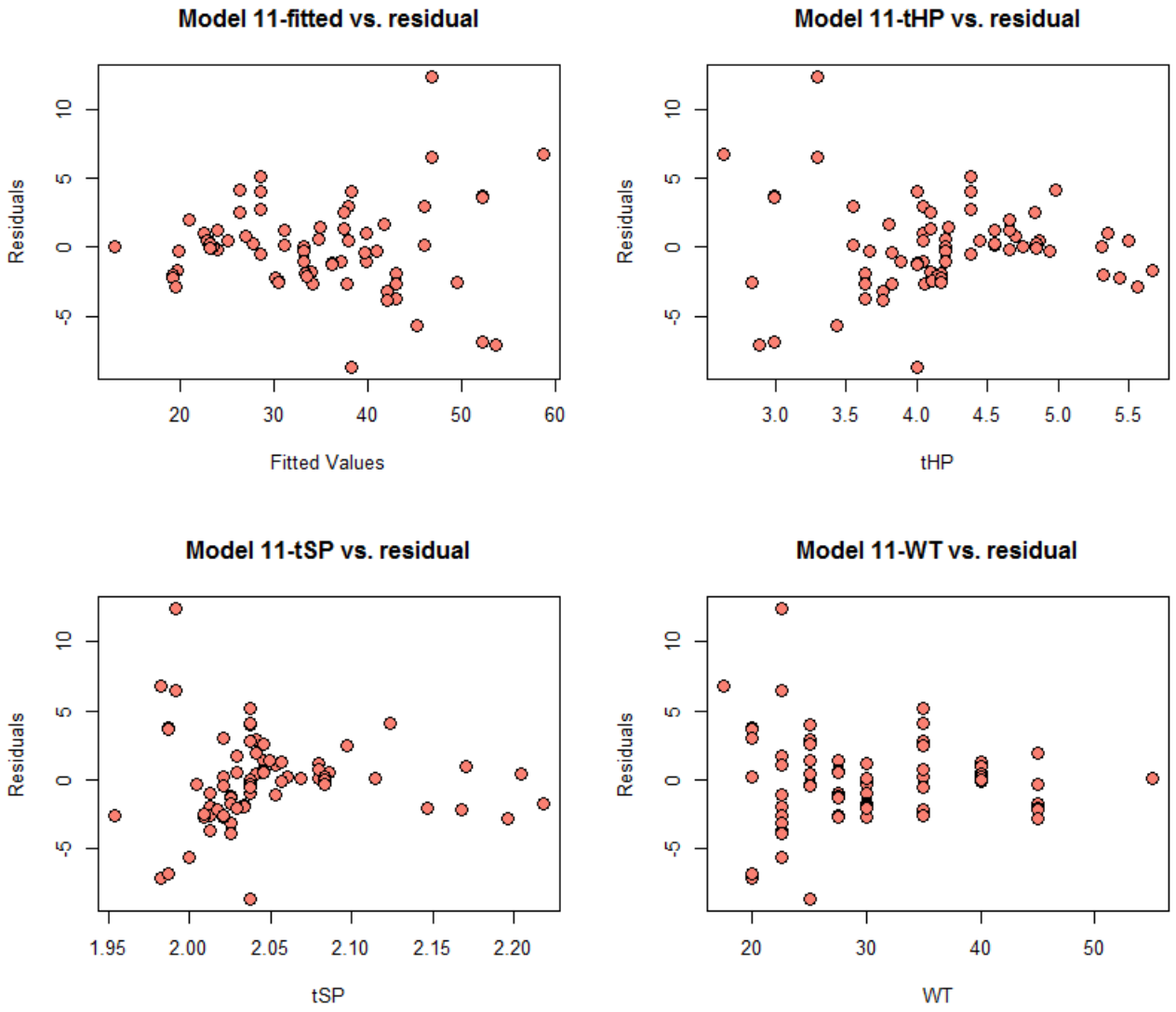

Figure 4 Residual Plots for Corrected Model 11

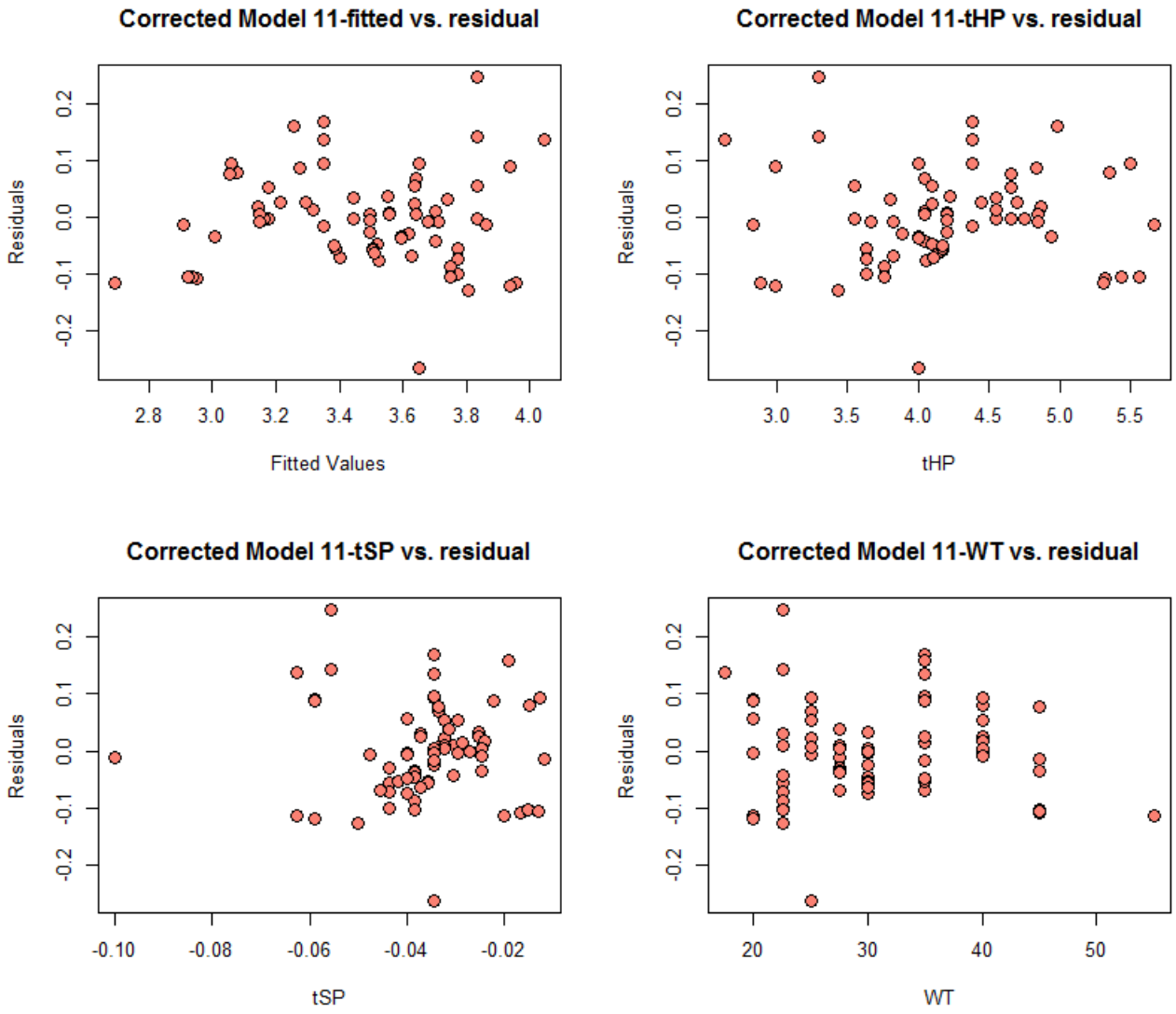

Figure 5 Outlier Diagnoses for Corrected Model 11

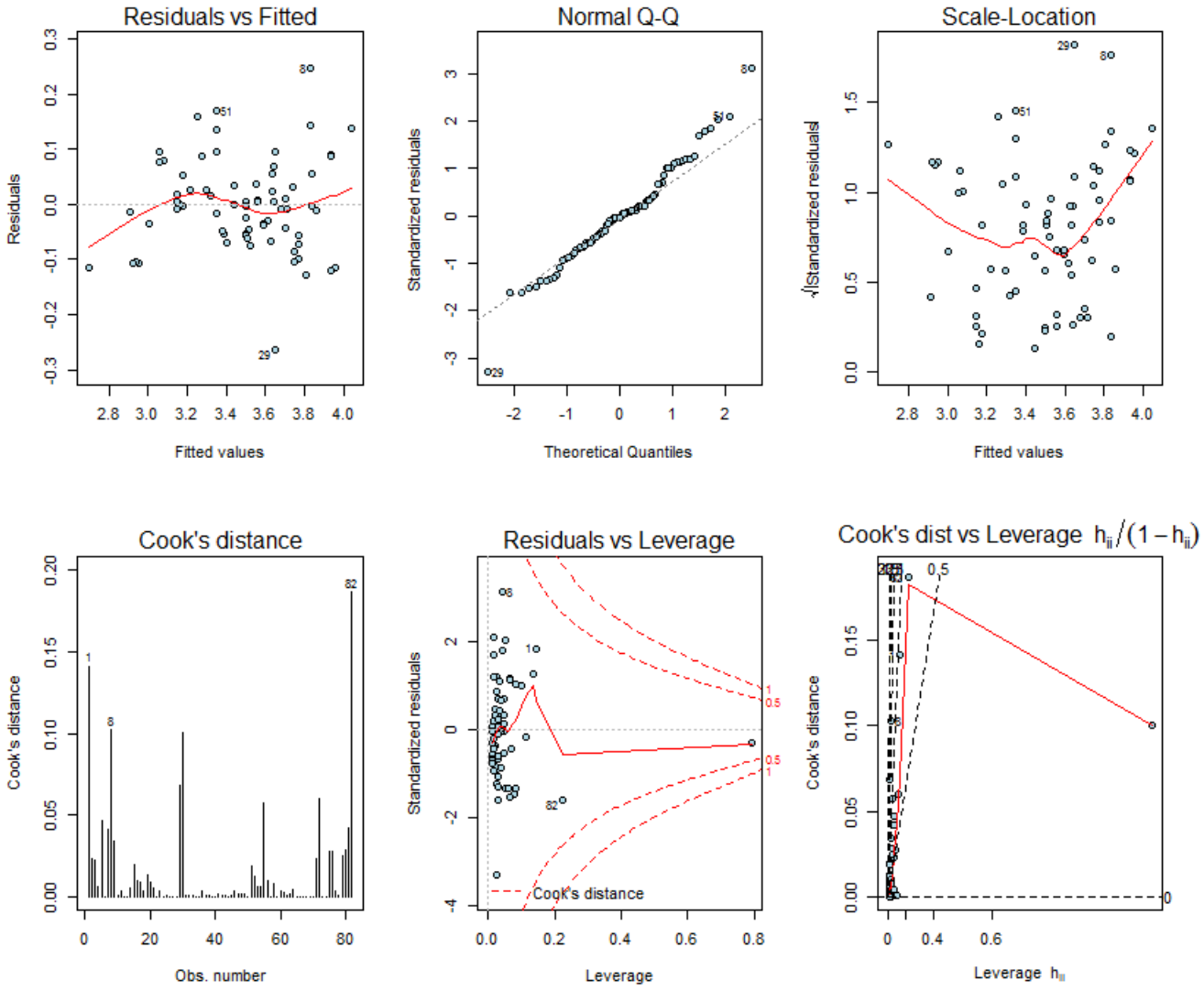

Figure 6 Residual Plots for Model 15

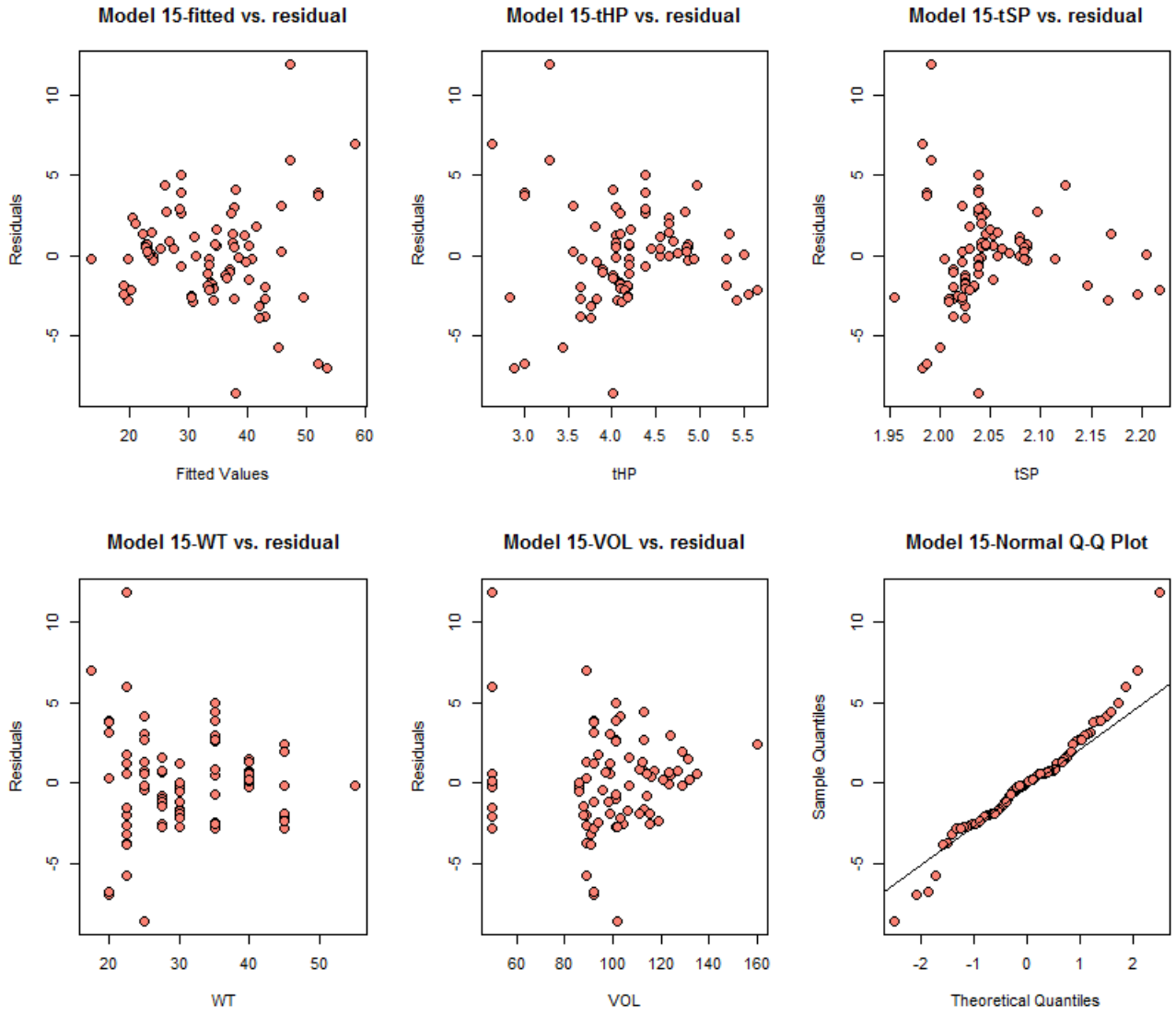

Figure 7 Residual Plots for Corrected Model 15

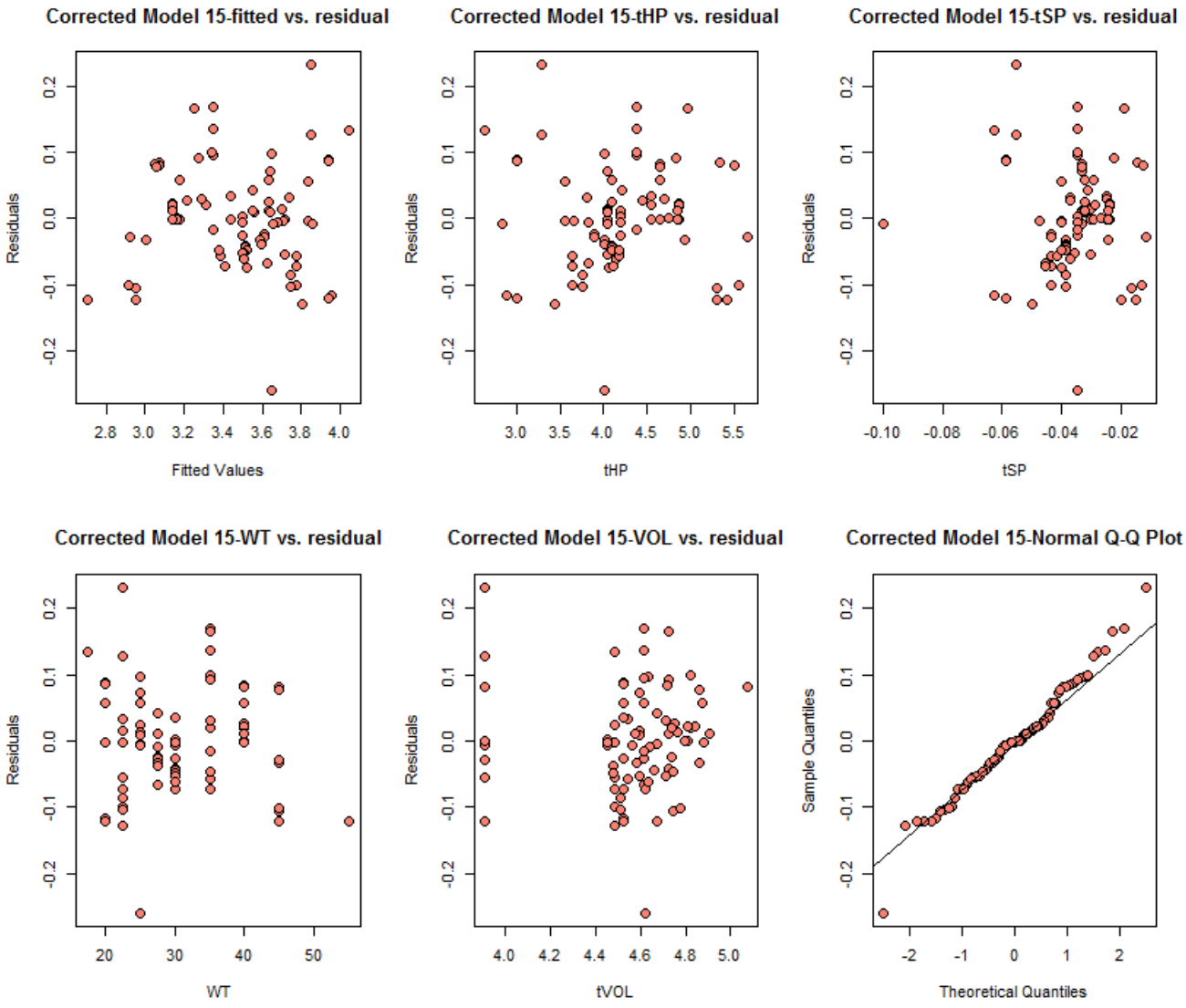

Figure 8 Outlier Diagnoses for Corrected Model 15

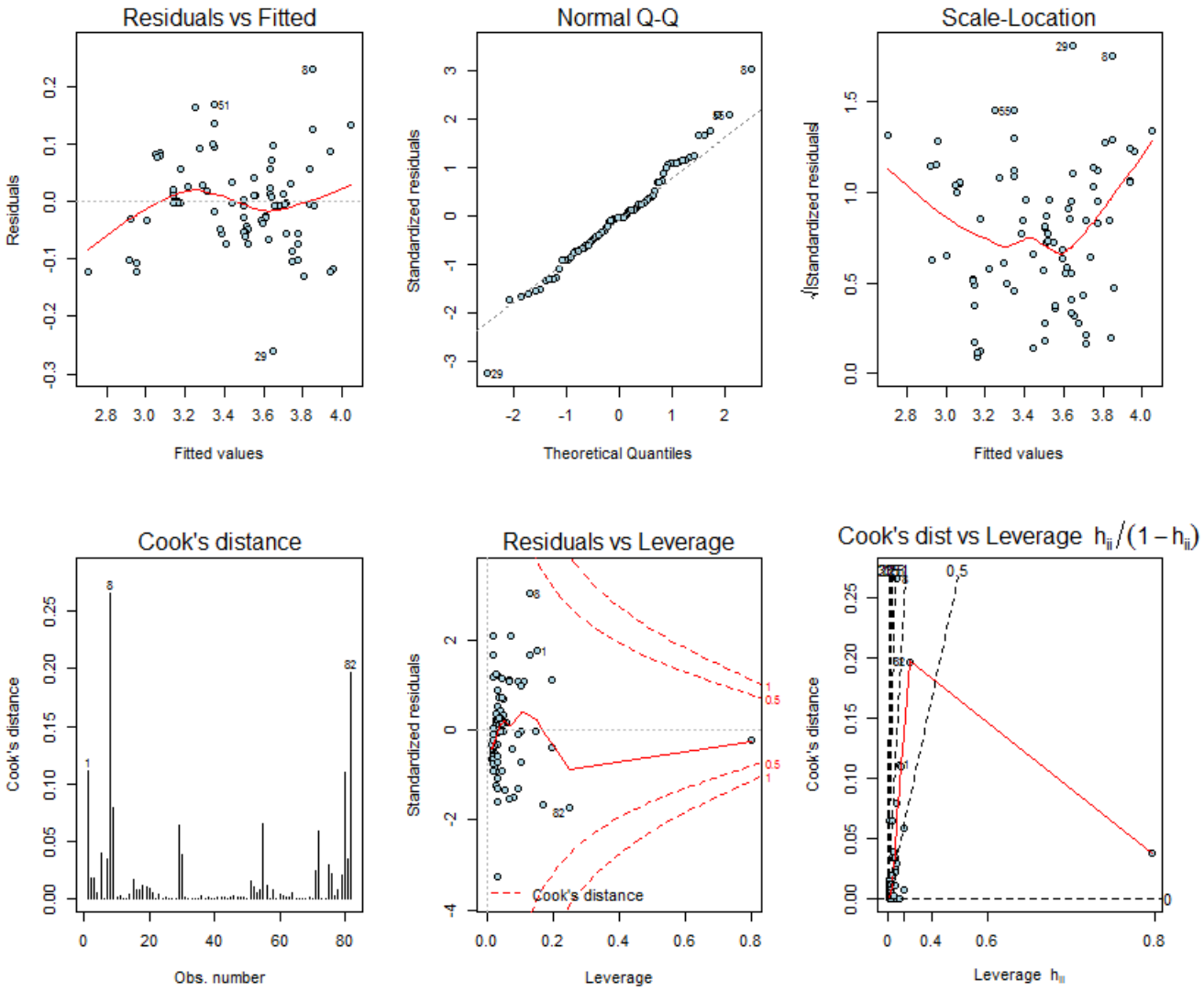

Figure 9 Residual Plots for Model 12

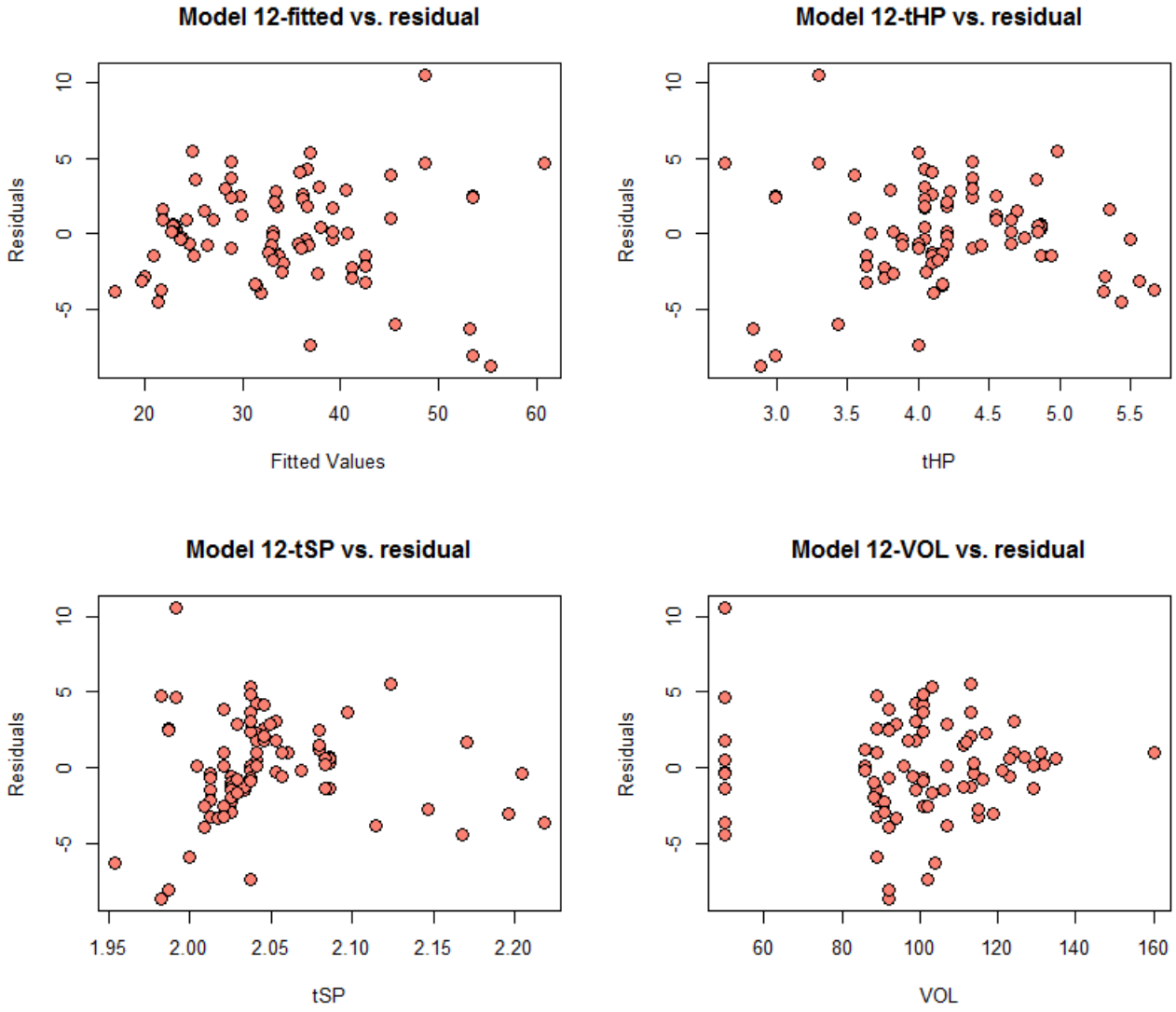

Figure 10 Residual Plots for Corrected Model 12

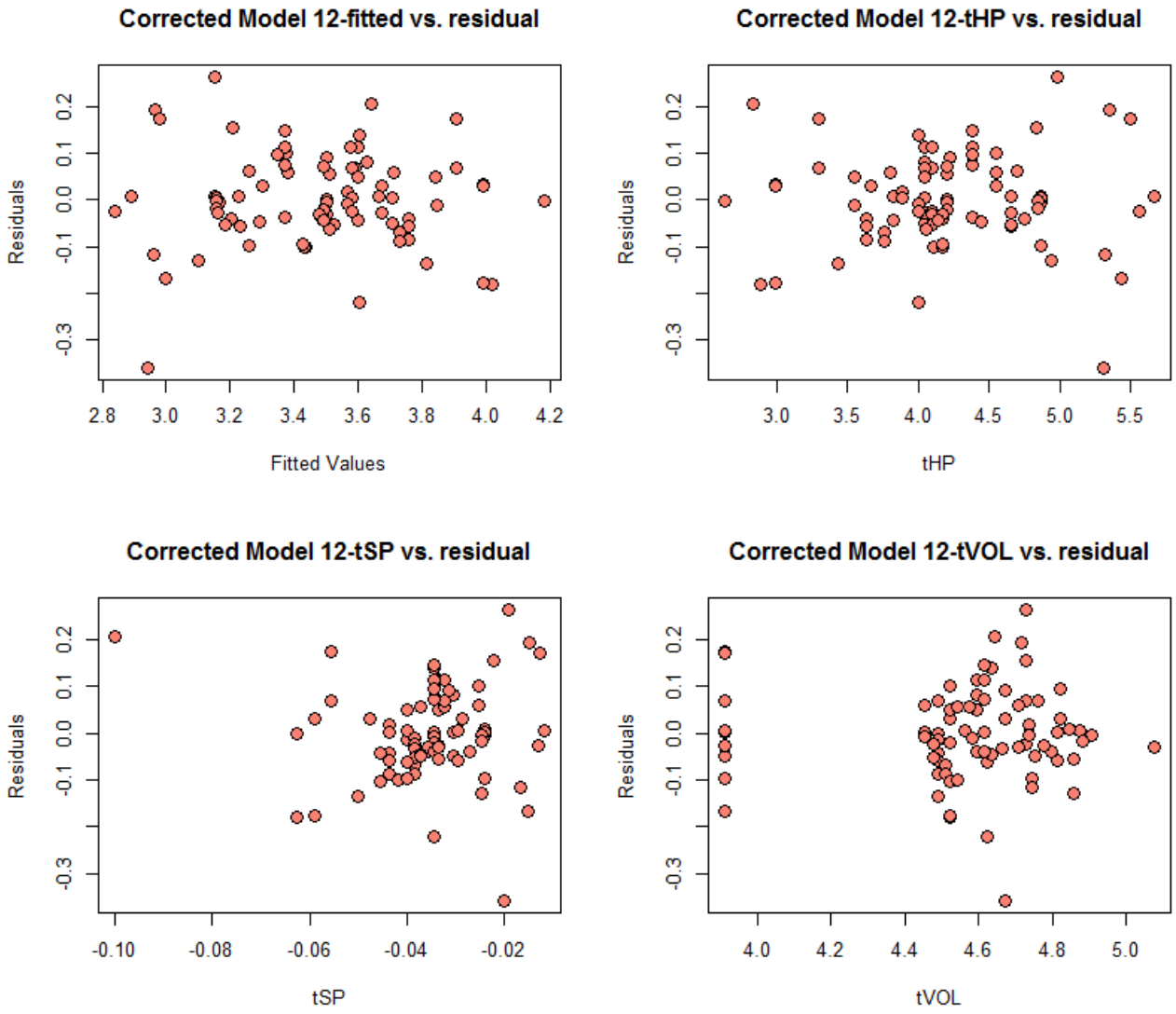

Figure 11 Outlier Diagnoses for Corrected Model 12