Supplementary Materials for the article: Robustness of Random Forests for Regression

Marie-Hélène Roy* Denis Larocque†

*Department of Management Sciences, HEC Montréal, 3000 chemin de la Côte–Sainte–Catherine, Montréal (Québec), Canada, H3T 2A7.

†Corresponding author. Department of Management Sciences, HEC Montréal, 3000 chemin de la Côte–Sainte–Catherine, Montréal (Québec), Canada, H3T 2A7.

1. SIMULATION STUDY WITH ARTIFICIAL DATA: THE CASE OF CONTAMINATED TEST DATA

In Section 3 of the main article, only the training data set is contaminated. In this section, we provide the results when the test data sets are also contaminated. They are summarized in Figure 1. In Section 3, since the true population was the same regardless of the proportion of contamination p, we used the % increase with respect to the non–contaminated case as the criterion. The goal was to investigate the impact of contamination (in the training data) to predict the same population. Now, the population varies with p and it is not really appropriate to use this criterion. Instead, we use the % increase in PMSE with respect to the best performer among the seven methods. Namely, if PMSE_i is the PMSE of method i ($i \in \{$RF, RFM, RFMR, QRF, RFLAD, RFLADM, RFLADWM$\}$) for a given configuration (i.e. choices of DGP, m, type of contamination and p) obtained as the average PMSE over 100 simulation runs, then the % increase in PMSE of method k with respect to the best performer is

$$100\frac{(\text{PMSE}_k - \min_i\{\text{PMSE}_i\})}{\min_i\{\text{PMSE}_i\}}.$$

Noting that the Y-axis scale does not go very high (only up to 12 % in two plots), we can see that all methods have a very similar performance across all configurations. In fact, this time, we are not really studying the impact of contamination but rather the relative performance of the seven methods for populations that contains some extreme data points. No clear patterns emerge and all methods perform relatively well compared to the others. Even the original RF does well, unlike the case where only the training data are contaminated.

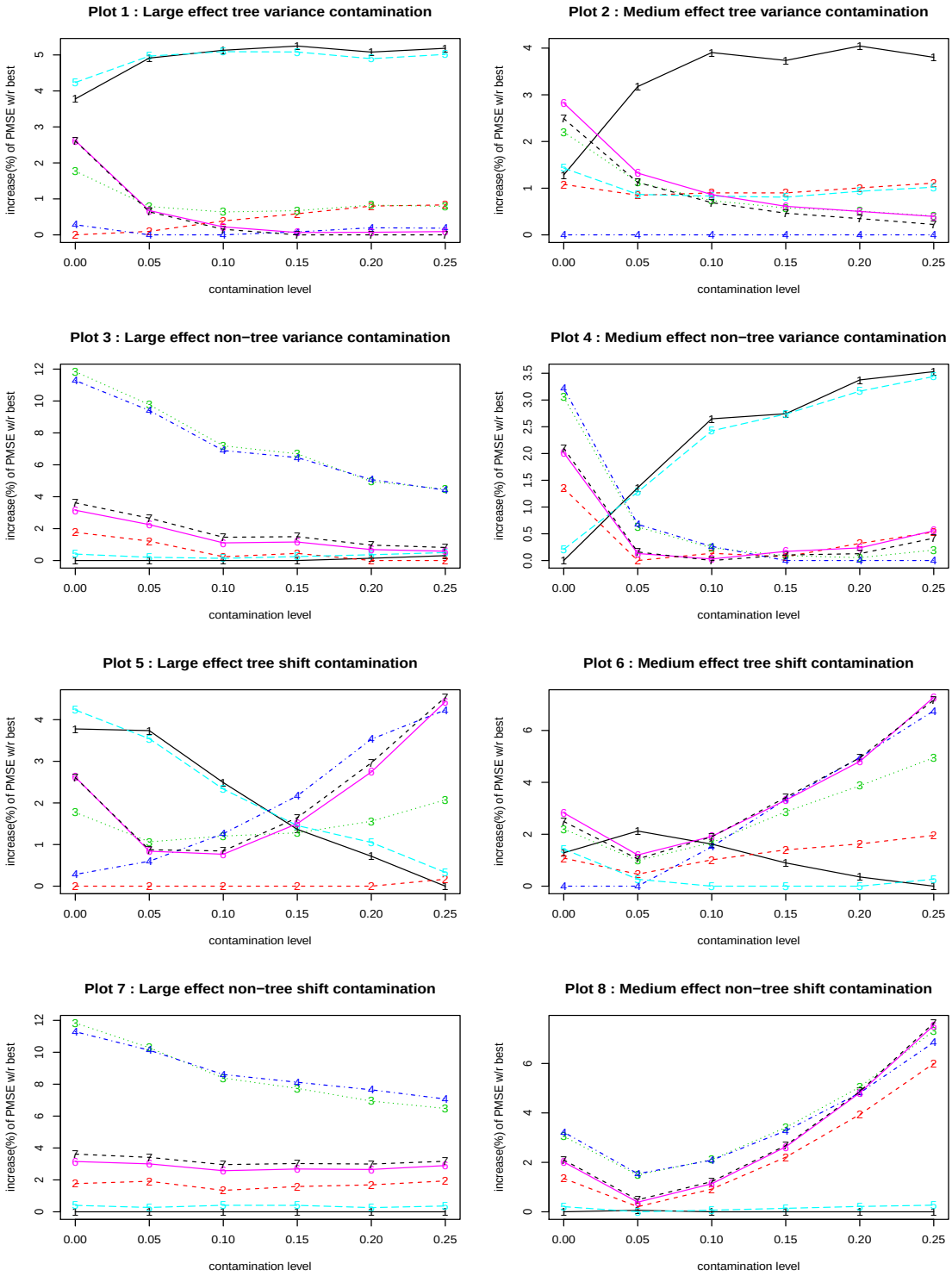

Figure 1: Behavior of the % increase in PMSE of each method when compared to the best performer for the same scenario as a function of the proportion of contamination. The lines are indexed in the following manner: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

2. REAL DATA SETS: USING PMSE AS THE CRITERION

In Section 4 of the main article, the performance of the seven methods are compared using the MAPE criterion. The rationale is that using a more robust evaluation criterion provides protection against possible outliers in the real data sets which may distort the comparisons. For completeness, the results with the PMSE criterion are presented here. Table 1 and Figure 2 present the results. This time, the original RF has the lowest median (1.38%) % increase in PMSE with respect to the best performer. However, its performance is not as stable across the data sets compared to RFM, RFLADM and RFLADWM. These three methods have the lowest maximum % increase and the lowest standard deviation. They also have the three smallest mean % increase among all seven methods. As for the case with the MAPE criterion, the performance of QRF is somewhat disappointing.

Table 1: Results with the real data sets. The upper part of the table presents the raw PMSE results and the lower part presents the % increase in the value of the criterion of each method when compared to the best performer for the same data set.

	PMSE						
Data set	RF	RFM	RFRM	QRF	RFLAD	RFLADM	RFLADWM
AUTO	7.51	7.41	7.68	8.40	7.55	7.65	7.72
BIRT	213333	215211	218788	205691	210825	221923	227831
BCWI	1111	1211	1211	1184	1100	1208	1210
CHIL	0.215	0.194	0.246	0.373	0.220	0.198	0.190
CCRI	0.018	0.018	0.019	0.022	0.018	0.018	0.018
COMP	3405	3500	6177	6314	3513	3880	3679
CONC	27.5	25.5	27.0	30.2	27.2	26.5	26.5
HORS	3.98	3.48	3.68	3.97	3.93	3.52	3.54
HOUS	10.5	10.6	12.0	10.8	10.3	10.9	10.9
LONG	160	161	167	178	160	165	165
PULS	256	225	244	240	247	233	234
SERV	0.720	0.612	0.861	0.792	0.714	0.524	0.527
CSLU	2060	2114	2181	2333	2085	2189	2183
	% increase of PMSE w/r best						
Data set	RF	RFM	RFRM	QRF	RFLAD	RFLADM	RFLADWM
AUTO	1.38	0.00	3.70	13.32	1.90	3.22	4.21
BIRT	3.72	4.63	6.37	0.00	2.50	7.89	10.76
BCWI	0.95	10.06	10.05	7.57	0.00	9.78	9.98
CHIL	13.05	1.96	29.28	96.09	15.70	4.15	0.00
CCRI	0.00	0.79	4.73	20.53	0.55	1.48	1.64
COMP	0.00	2.77	81.39	85.44	3.16	13.94	8.05
CONC	7.91	0.00	5.74	18.30	6.71	3.76	3.80
HORS	14.37	0.00	5.73	14.07	13.07	1.22	1.70
HOUS	1.33	2.11	16.19	4.72	0.00	5.06	5.49
LONG	0.05	0.54	4.40	11.71	0.00	3.06	3.51
PULS	13.73	0.00	8.45	6.87	9.92	3.53	4.33
SERV	37.52	16.93	64.32	51.23	36.42	0.00	0.65
CSLU	0.00	2.60	5.86	13.23	1.19	6.25	5.97
mean	7.23	3.26	18.94	26.39	7.01	4.87	4.62
median	1.38	1.96	6.37	13.32	2.50	3.76	4.21
maximum	37.52	16.93	81.39	96.09	36.42	13.94	10.76
Standard deviation	10.70	4.96	25.14	31.19	10.29	3.82	3.38

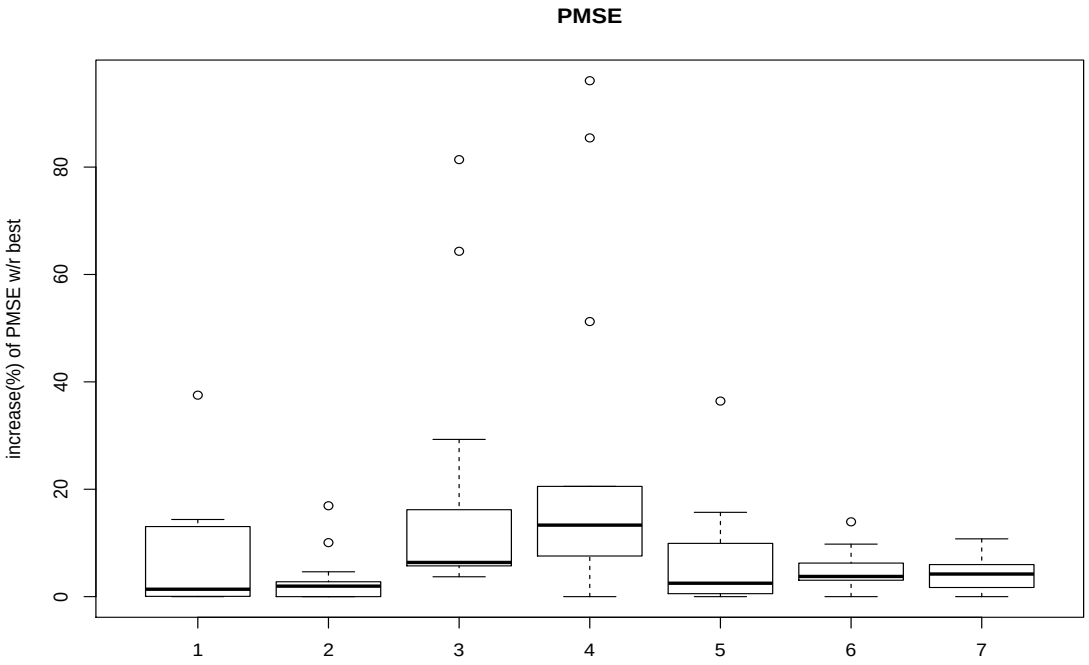

Figure 2: Results for the real data sets. Distribution of the % increase in the value of PMSE of each method when compared to the best performer for the same data set. The figure is labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

3. REAL DATA SETS: THE IMPACT OF ADDING CONTAMINATION

In Section 4 of the main article, we investigate the relative performance of the seven methods with 13 real data sets. The goal there is not to study directly the robustness aspects, since the type, amount and even what constitutes an outlier, in these data sets is unknown. Here, to complement the investigation with artificial data, we will study the robustness aspects by adding artificial contamination to the real data sets. Once again, two types of contamination are considered: variance and shift. Let σ_Y denote the standard deviation of the outcome for the data set. For the variance contamination, with probability p, a normal variate with mean 0 and standard deviation $5 * \sigma_Y$ is added to the outcome. For the shift contamination, with probability p, a normal variate with mean $5 * \sigma_Y$ and standard deviation σ_Y is added to the outcome. Only the training data set is subject to contamination in the cross–validation scheme and not the test part. Moreover, only $p = 15\%$ contamination is considered. The results are presented in Figure 3. Both % increase in MAPE with respect to the best performer and % increase in MAPE with respect to the non–contaminated case are reported. It is clear from these plots that RF and RFLAD, the only two methods using the mean to aggregate the individual trees, are a lot more affected by contamination than the others. As for artificial data in Section 3 of the main article, QRF offers a very good protection against contamination. The other methods using median type aggregation (RFM, RFRM, RFLADM and RFLADWM) are also substantially better than RF and RFLAD. Hence, once again, it seems that using a robust aggregation method is more beneficial than using only a robust splitting criterion (RFLAD).

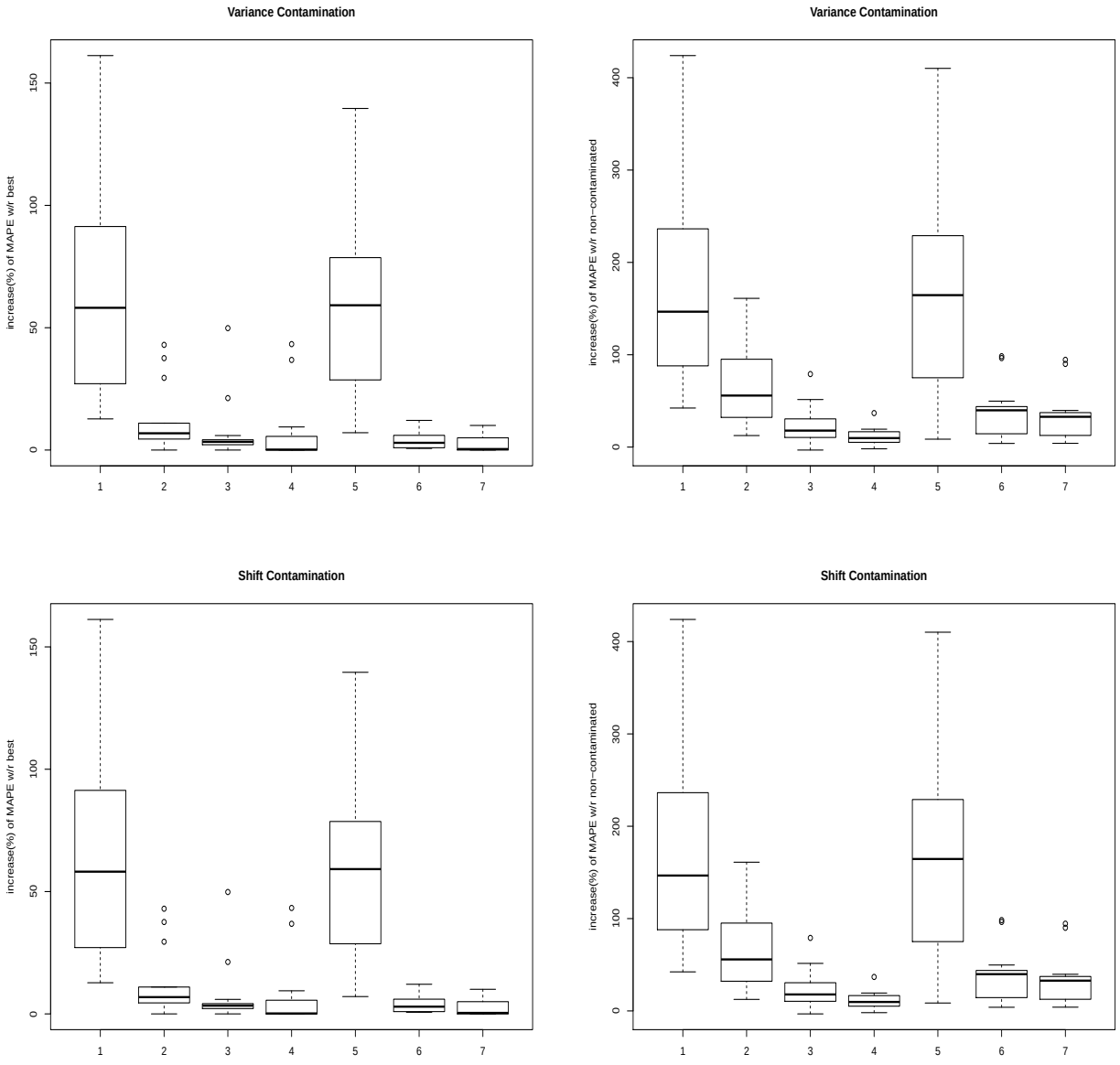

Figure 3: Results for the real data sets. The two upper plots present the results with the variance contamination and the two lower plots present the results with the shift contamination. The plots on the left present the distribution of the % increase in the value of MAPE of each method when compared to the best performer for the same data set and the plots on the right present the distribution of the % increase in the value of MAPE of each method when compared to the non-contaminated case. The figure is labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

4. REAL DATA SETS: A CLOSER LOOK AT THE RESULTS THROUGH RESIDUALS

Even if QRF is seen to offer quite a good protection against the type of contamination considered in this article, as evidenced by the results in Section 3 of the main article and of the preceding Section in the Supplementary Materials, it is slightly worse than the others with the real data sets (the original data sets without added contamination); see Section 4 of the main article. In order to investigate this we provide a closer look at each data set in this Section. There is one figure for each data set. Each figure contains four plots. The upper left plot is an histogram of the outcome. The other three plots involve residuals obtained from the same 30 repetitions of 10–fold cross–validation that produced the results in Section 4 on the main article. The upper right plot contains box–plots of the residuals for all methods. The lower two plots are used to investigate the performance of QRF. Since RFLADWM is the most stable method across all data sets, it is used as the point of comparison to try to find out where QRF is making more errors. Let $\hat{e}_{QRF}$ and $\hat{e}_{RFLADWM}$ denote the residuals for QRF and RFLADWM, respectively. The lower left plot is a scatter plot of $|\hat{e}_{QRF}|$ - $|\hat{e}_{RFLADWM}|$ as a function of the outcome Y. When $|\hat{e}_{QRF}|$ - $|\hat{e}_{RFLADWM}|$ is greater than 0, QRF makes a larger error (in absolute value) than RFLADWM, and vice–versa. Thus, this plot allows us to see where (in the Y space) QRF is making larger errors, as compared to the good and stable method RFLADWM. The lower right plot is simply a plot of $\hat{e}_{QRF}$ as a function of $\hat{e}_{RFLADWM}$. A loess smoother is added to each lower plots.

Let's look first at the Breast Cancer data set (Figure 5) where QRF and RFLADWM have a similar performance (MAPE of 28.8 and 29.0, respectively). We can see that the outcome has a right–skewed distribution without any extreme values (upper left plot). The box–plots of the residuals are very similar for the seven methods (upper right plot). The lower left plot does not show any strong tendency. RFLADWM is making a little more errors in the lower part of the Y distribution and the opposite

happens in the upper part. The lower right plot shows that the residuals of both methods are strongly correlated.

At the other extreme, the Childhood data set (Figure 8) is one where RFLADWM is the best method and QRF the worse one with a % increase in MAPE of 38.3% (compared to the best). Once again, the outcome is slightly right–skewed to the right and no extreme values are apparent. The box–plots of the residuals of QRF show more variability (a larger box) and more extreme residuals, compared to the other methods. The lower left plot is more informative. It shows that QRF makes systematically larger errors in the lower and upper part of the Y distribution. It is striking that QRF has a lot more problems predicting the high values of Y compared to RFLADWM.

Another data set where QRF is not doing well compared to RFLADWM is Servo (Figure 16). This time Y has a highly right–skewed distribution. Once again, QRF makes a lot more error in predicting the high values of Y.

The other graphs show different behaviors and it is not obvious to find characteristics of the data sets that would let us anticipate a good or bad performance of QRF. As advocated in the article, the best approach is to try several methods and select the best one with an out–of–sample or cross–validation scheme.

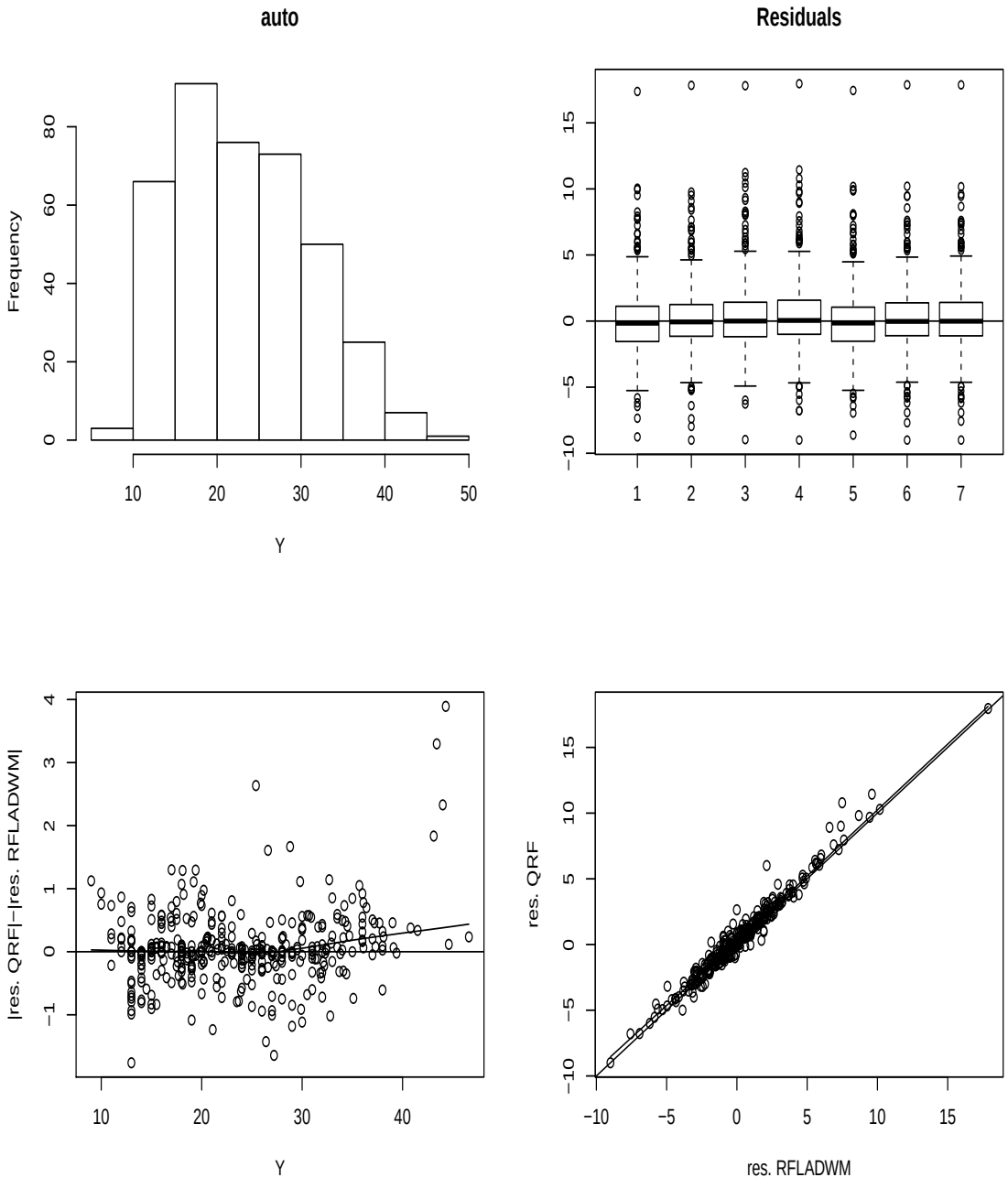

Figure 4: AUTO data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

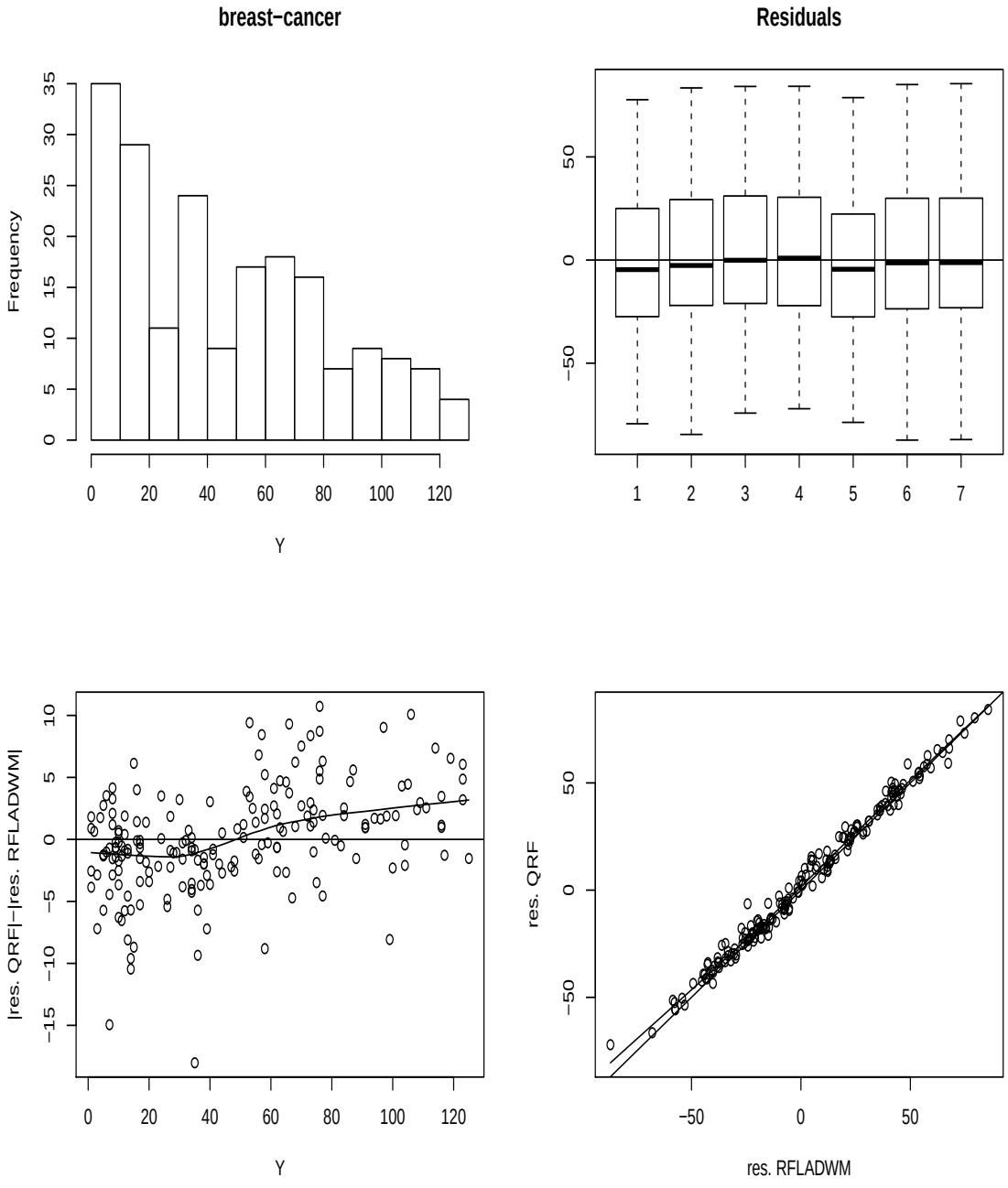

Figure 5: BCWI data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

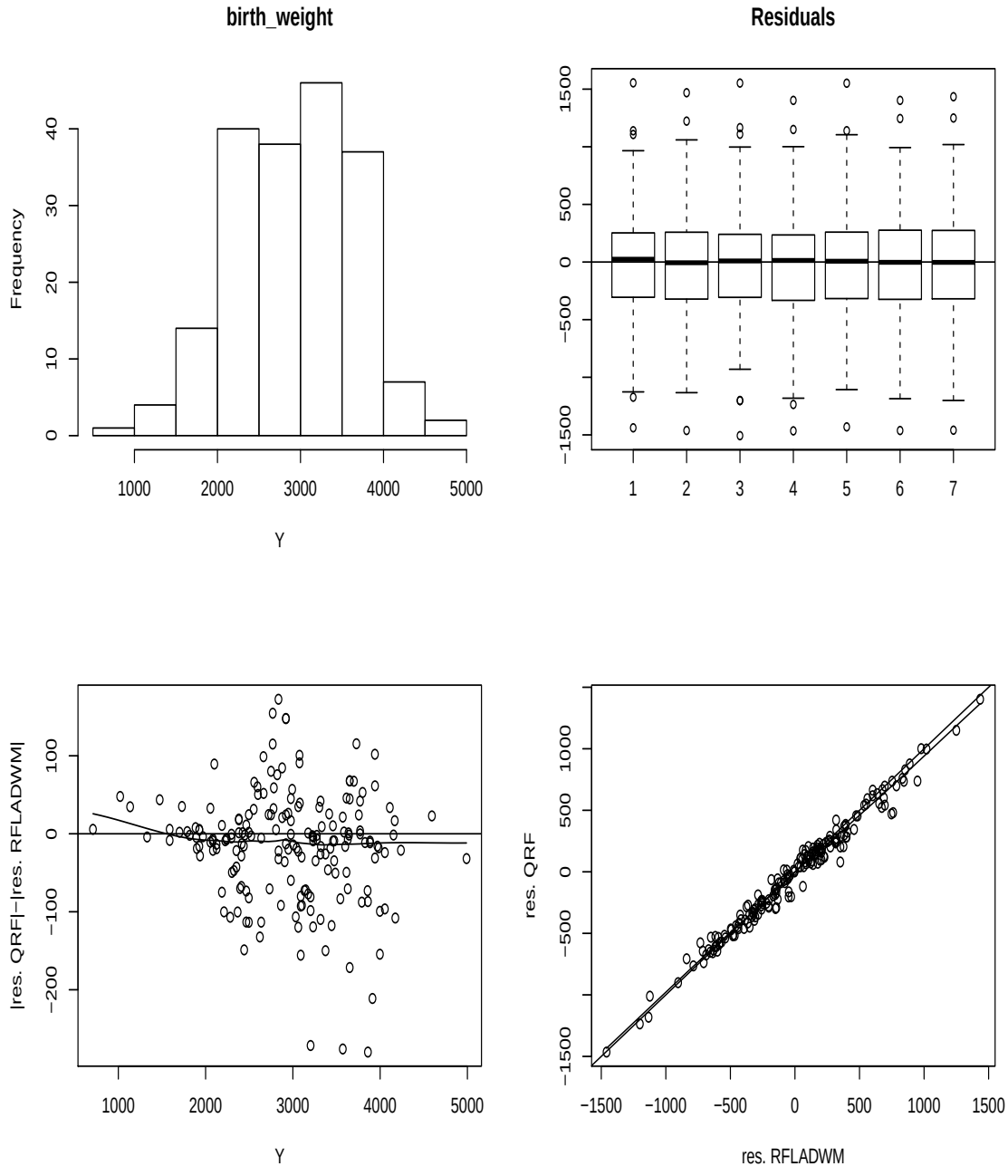

Figure 6: BIRT data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

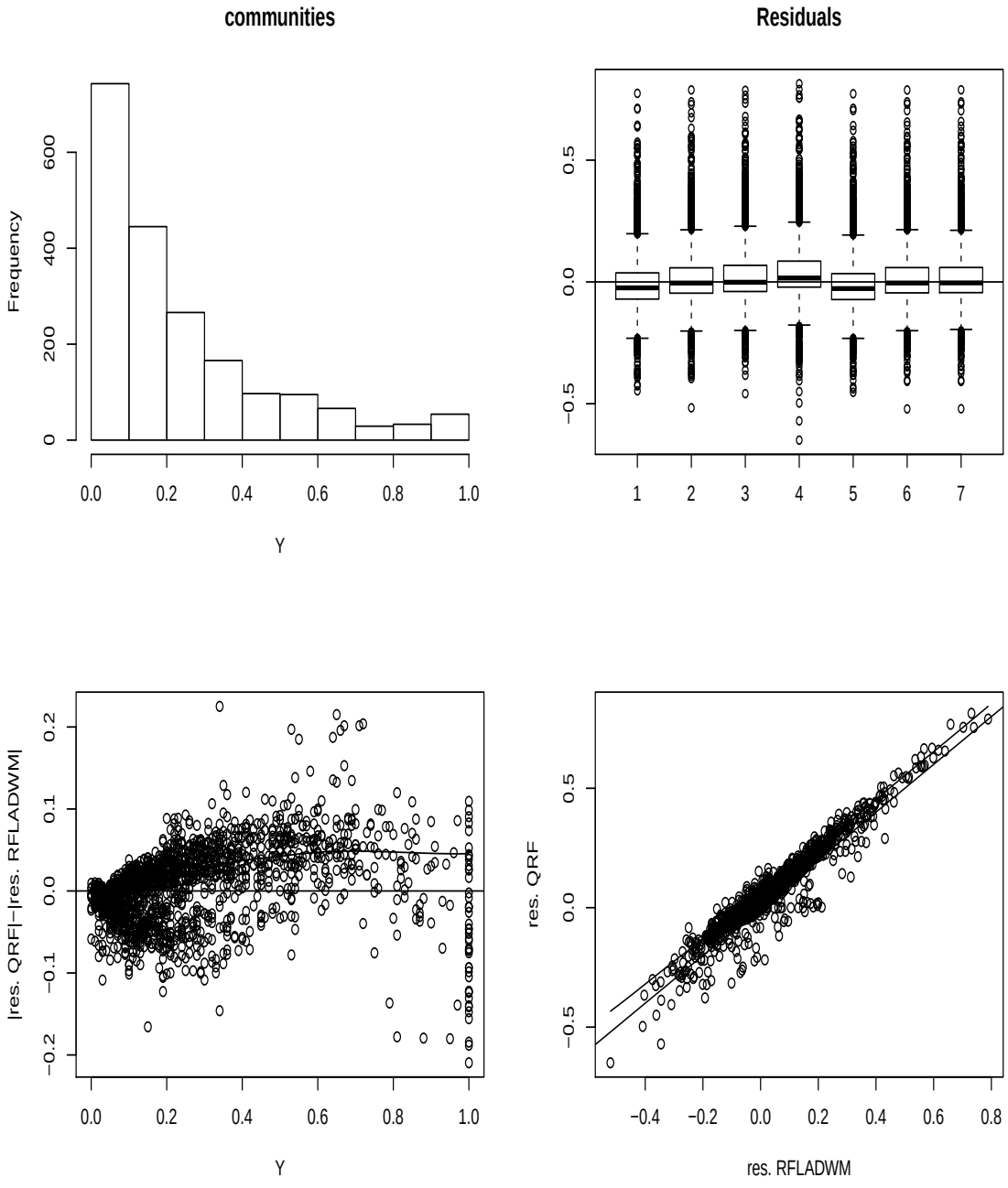

Figure 7: CCRI data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

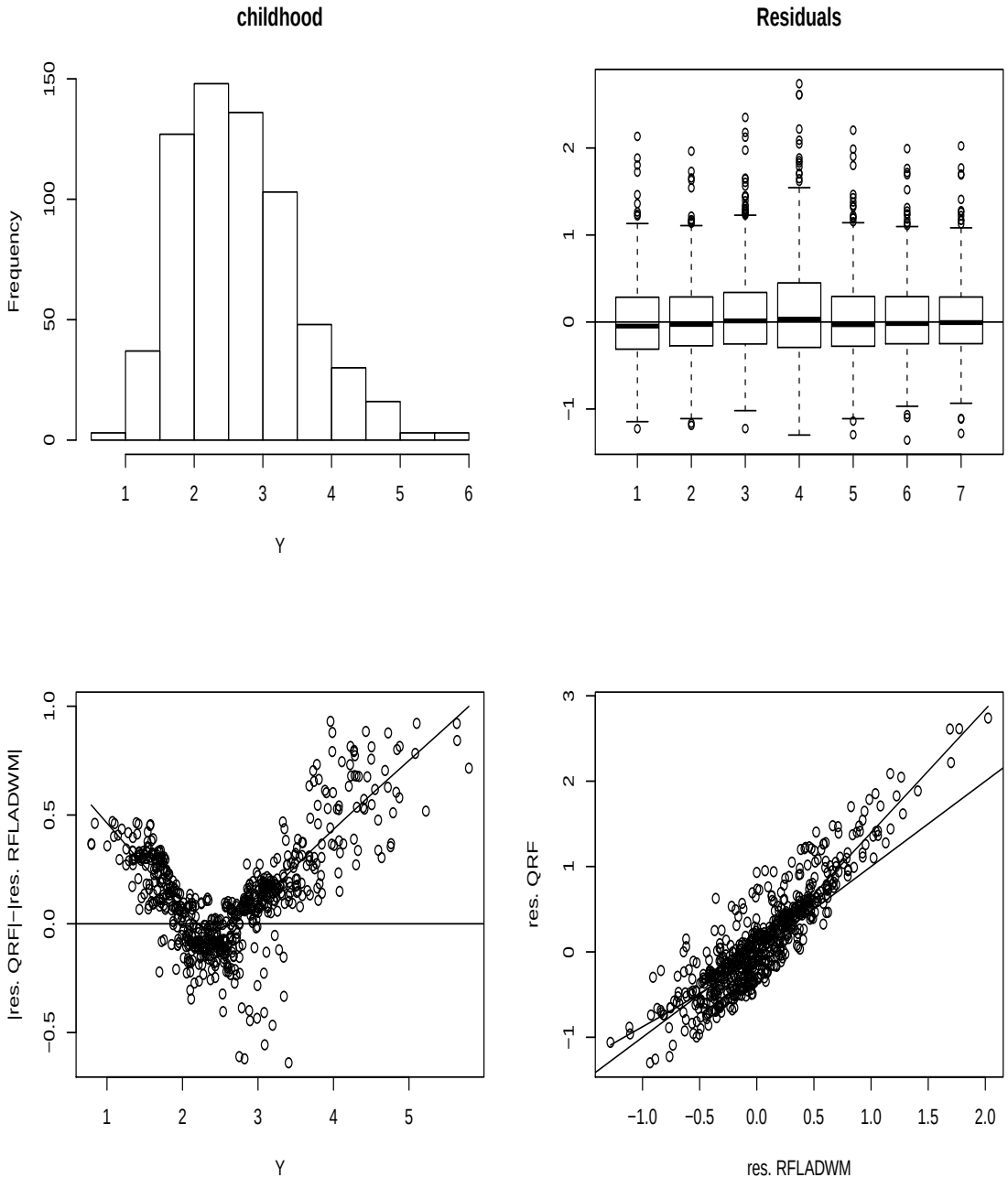

Figure 8: CHIL data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

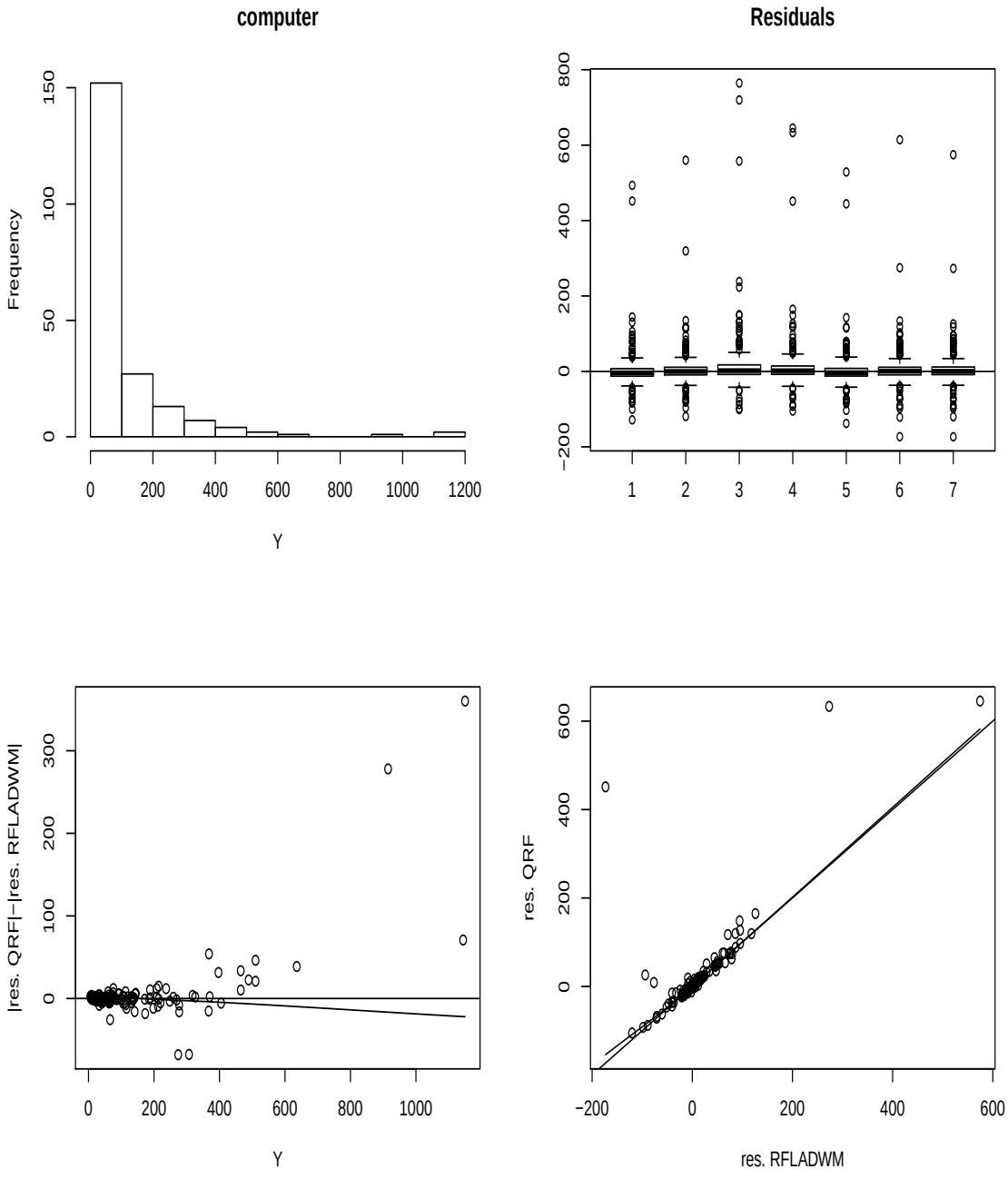

Figure 9: COMP data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

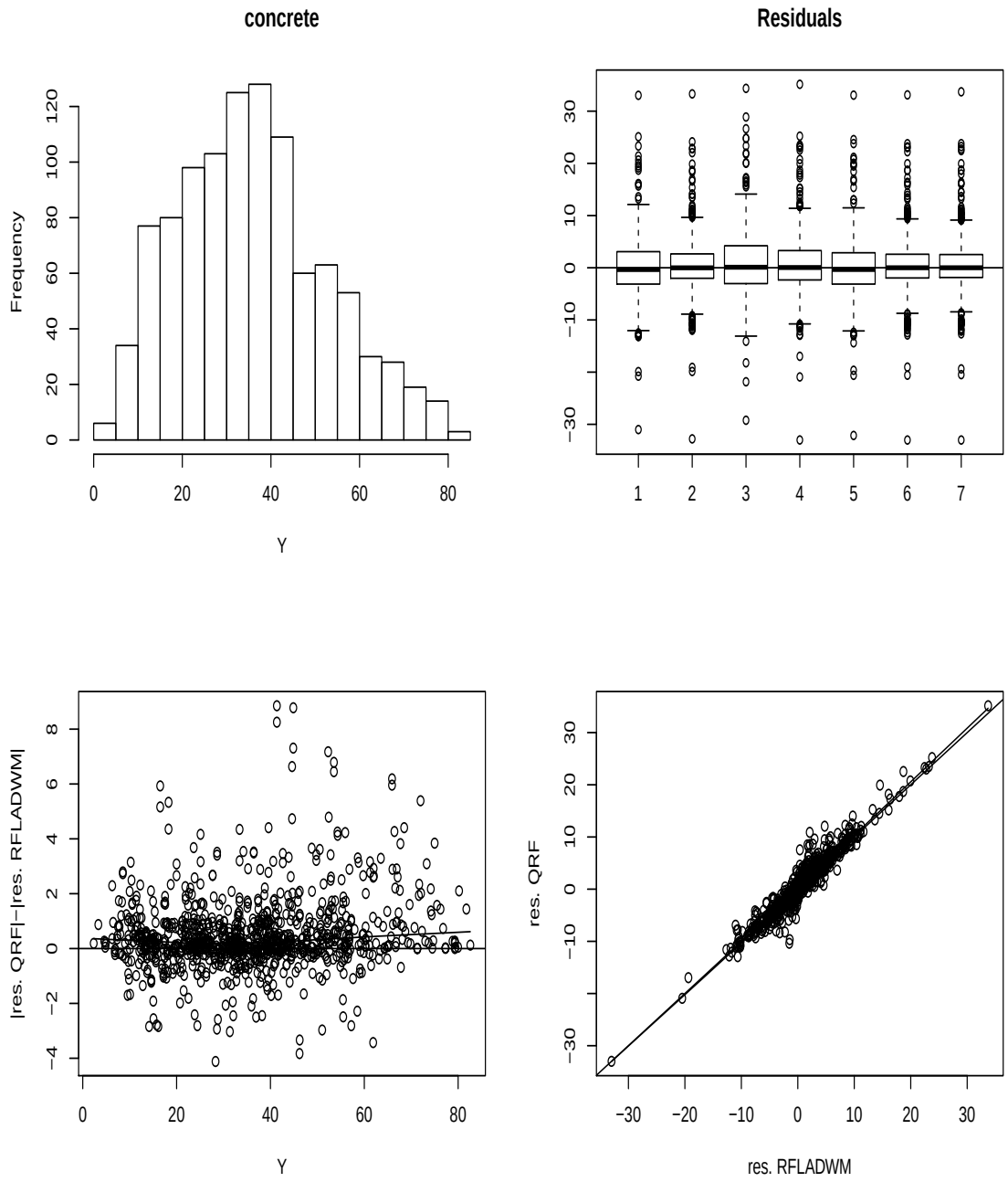

Figure 10: CONC data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

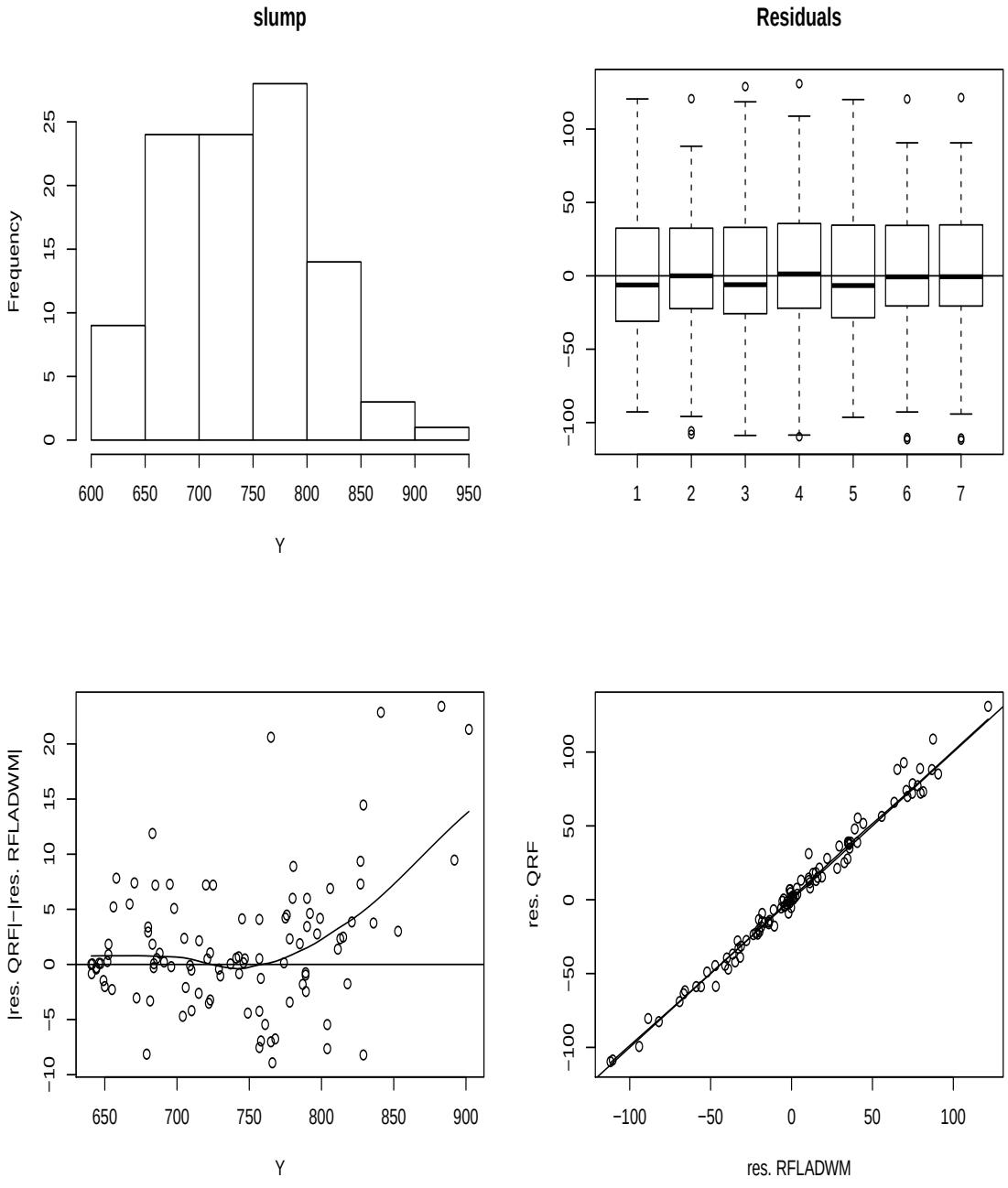

Figure 11: CSLU data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

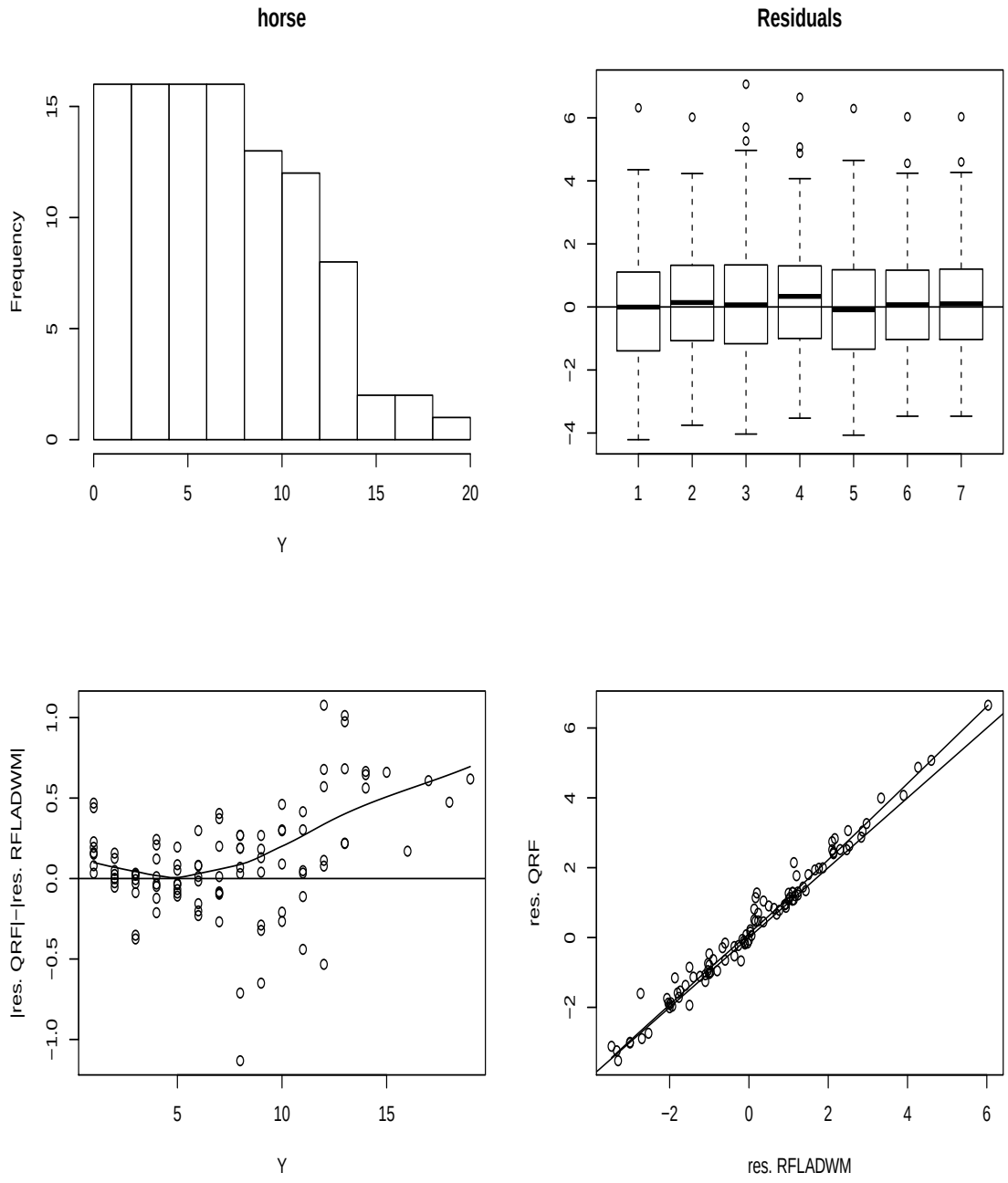

Figure 12: HORS data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

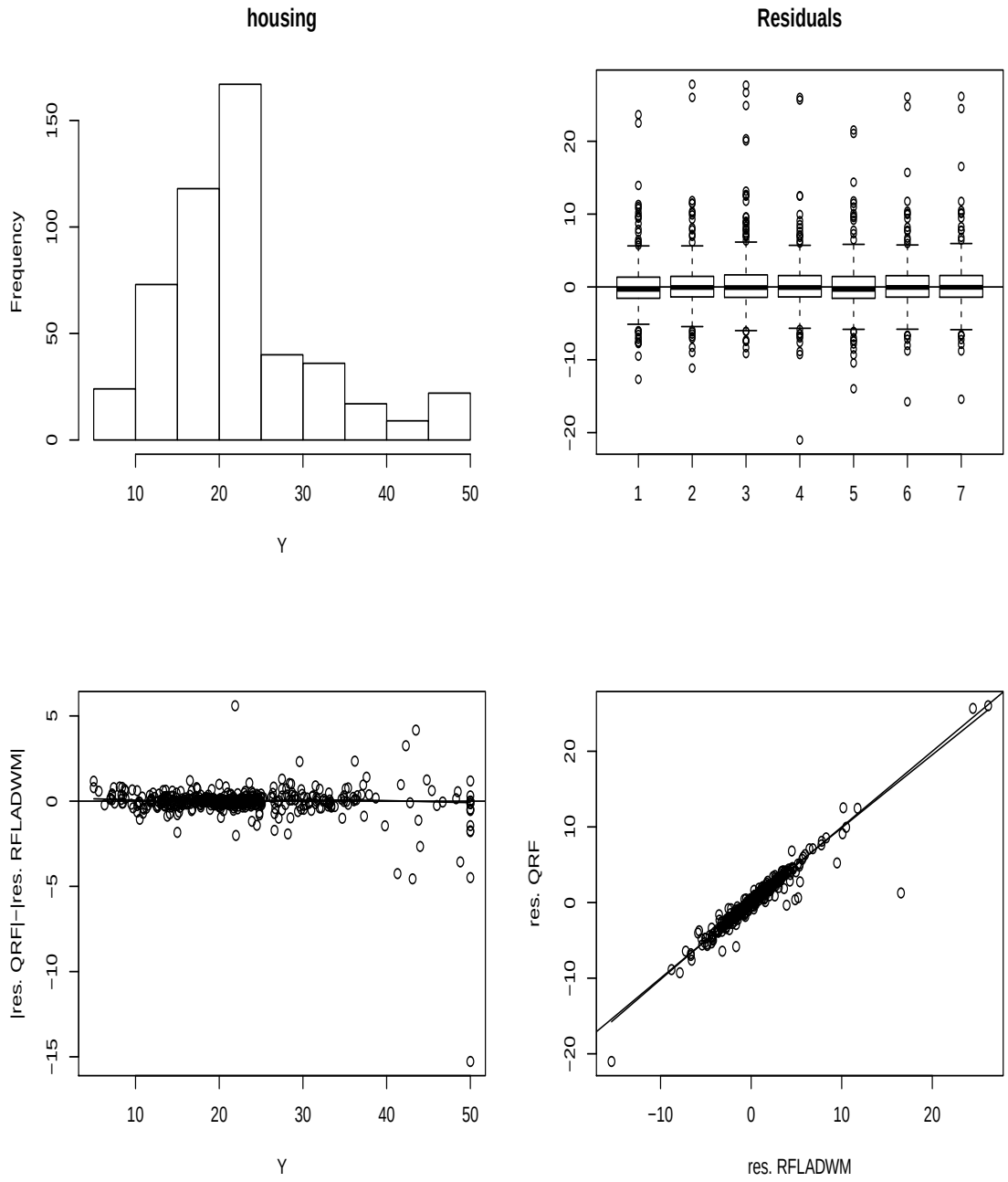

Figure 13: HOUS data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

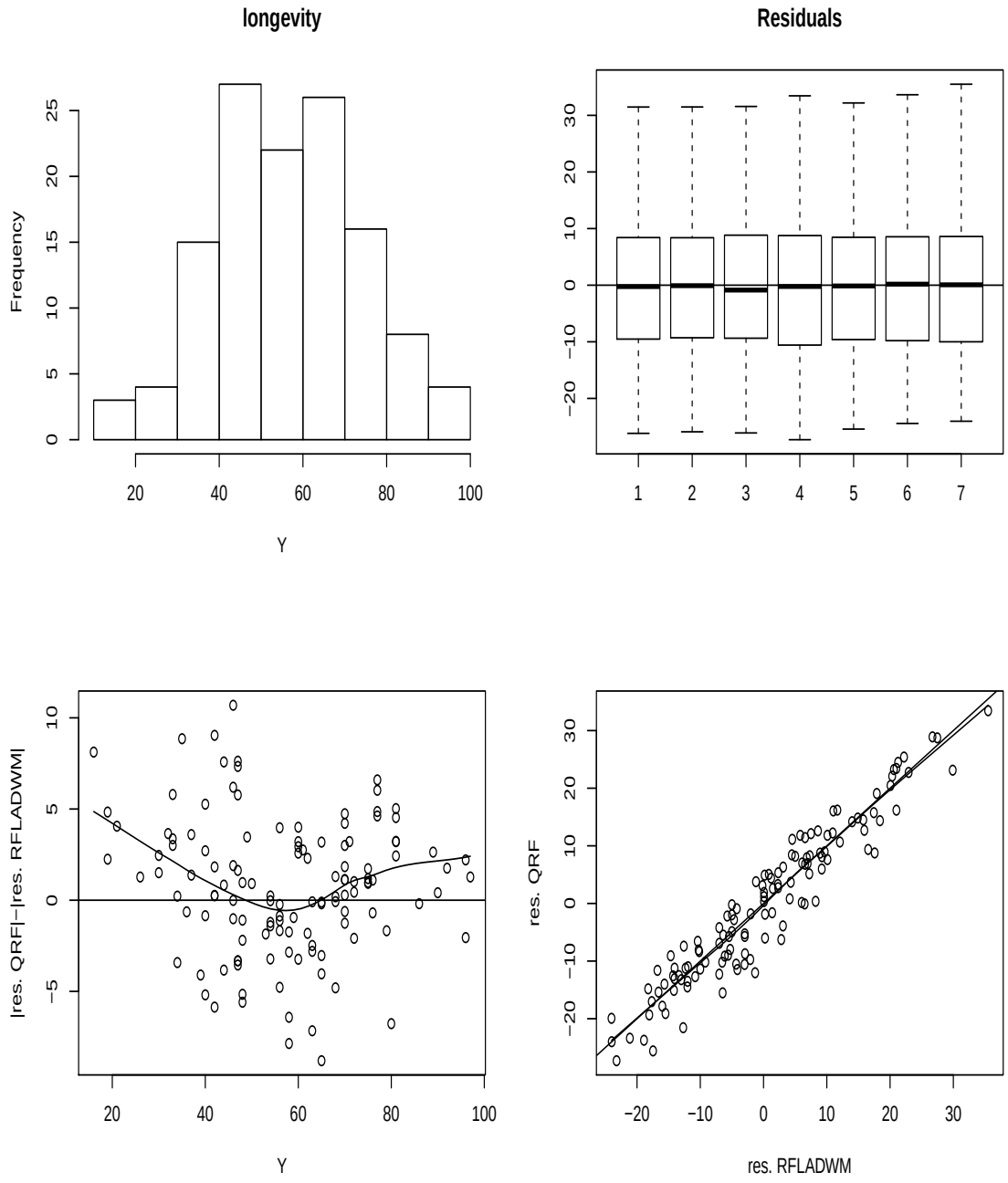

Figure 14: LONG data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

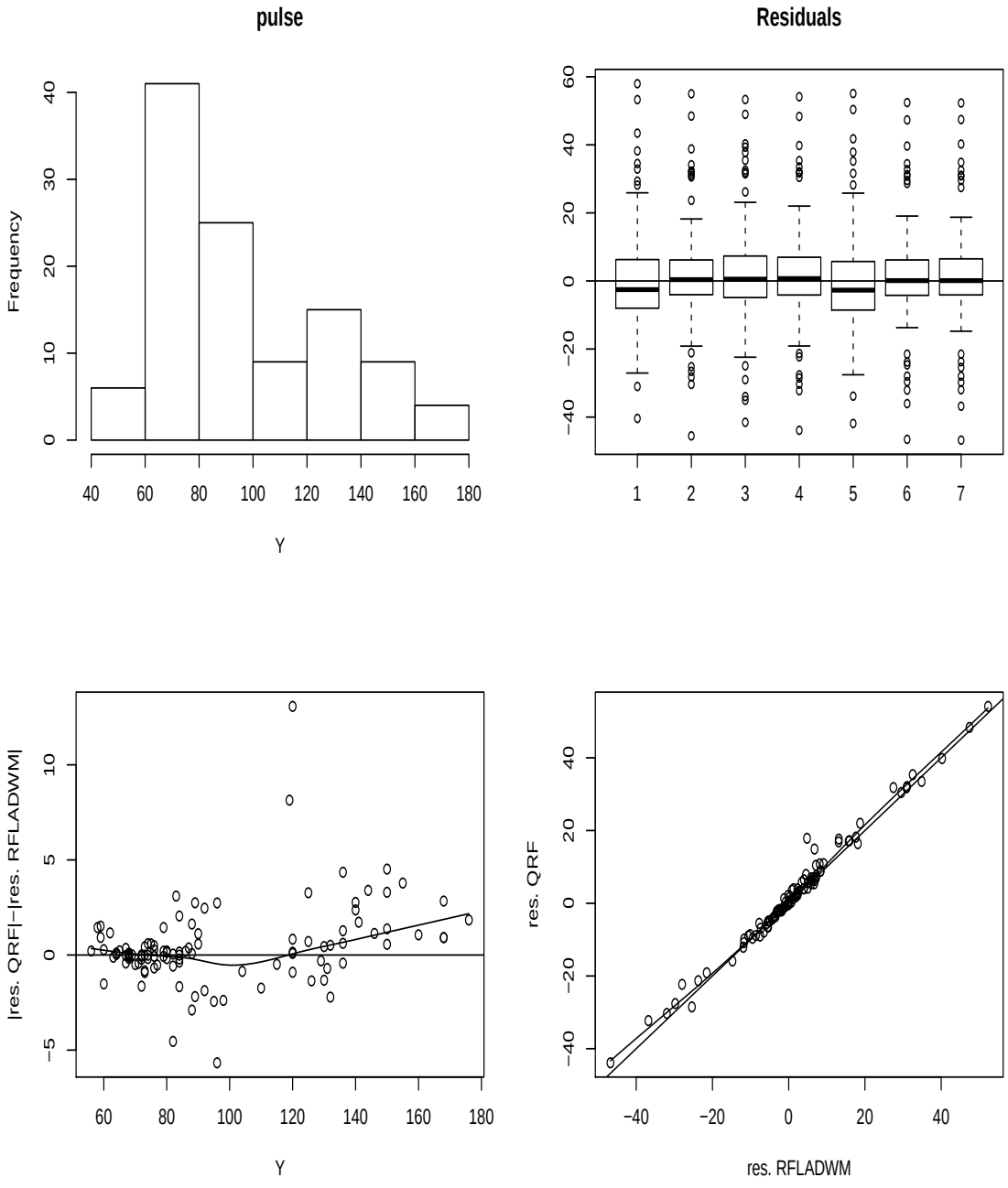

Figure 15: PULS data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.

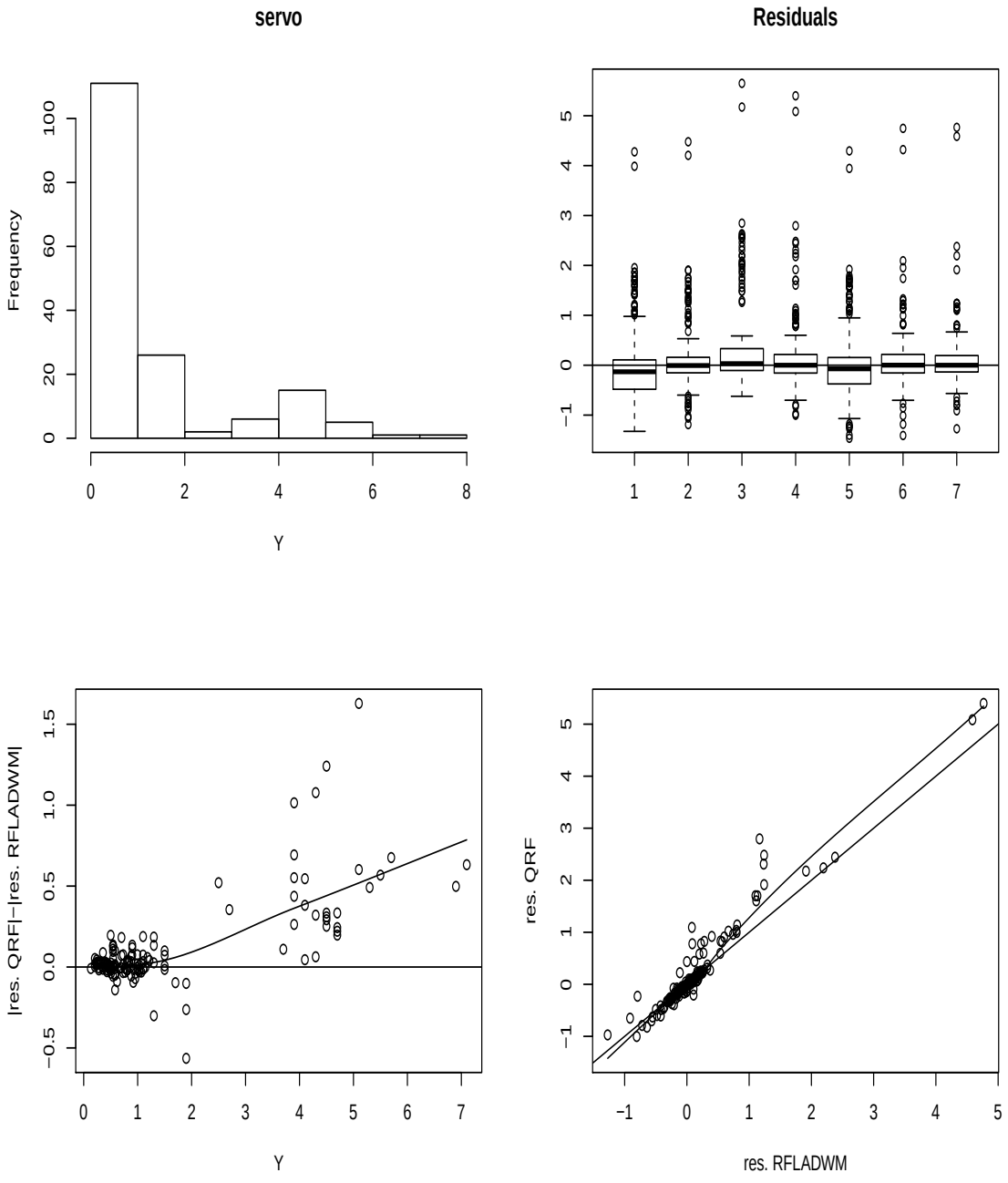

Figure 16: SERV data set. In the upper right plot, the box–plots are labeled as follows: 1=RF, 2=RFM, 3=RFRM, 4=QRF, 5=RFLAD, 6=RFLADM, 7=RFLADWM.