AMA3602 Applied Linear Models

Department of Applied Mathematics

Lecturer : Dr. Catherine Liu

02/2024

Ch3 I---Residual Diagnostics

Introduction

- Major assumptions of S/M-LR:
  1. The relationship between the response y and the regressors x's are linear, at least approximately;
  2. The error term ε has zero mean, i.e. $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$;
  3. The error term ε has constant variance σ^2, i.e. $Var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$;
  4. The errors are uncorrelated, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$ $(i \neq j)$, $i, j = 1, \ldots, n$;
  5. The errors are normally distributed, i.e. $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$.

  Remark:
  - ⋆ Assumptions 4 and 5 together imply that the errors are independent random variables;
  - ⋆ Assumption 5 is required for hypothesis testing and interval estimation.
- Question to be solved: Validity of aforementioned assumptions.
  - ⋆ The types of model inadequacies discussed here have potentially serious consequences;
  - ⋆ Gross violations of the assumptions may yield an unstable model in the sense that a different sample could lead to a totally different model with opposite conclusions;
  - ⋆ We usually cannot detect departures from the underlying assumptions by examination of the standard summary statistics, such as the t or F statistics, or R^2. These are "global" model properties, and as such they do not ensure model adequacy.

  In this chapter, we present several methods for diagnosing violations of the basic regression assumptions. These diagnostic methods are primarily based on study of the model **residuals**.

Residual Analysis

- Definition of residuals
- Methods for scaling residuals
- Residual plots
- Other residual plotting and analysis methods

Review of residuals

- **Residuals**: $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$.
  - ⋆ A residual may be viewed as the **deviation** between the **data** and the **fit**;
  - ⋆ It is also a measure of the variability in the response variable y not explained by the regression model;
  - ⋆ It is also convenient to think of the residuals as the realized or observed values of the model errors.

  Thus, any departures from the assumptions on the errors should show up in the residuals.
- Analysis of the residuals is an effective way to discover several types of model inadequacies.
- **Plotting residuals** is a very effective way to investigate how well the regression model fits the data and to check the aforementioned assumptionss
- The residuals have several important properties:
  - ⋆ $E(e_i) = 0$, $i = 1, \ldots, n$;
  - ⋆ $\frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res}$, unbiased estimator;
  - ⋆ The residuals are not independent, however, as the n residuals have only $n - p$ degrees of freedom associated with them. This nonindependence of the residuals has little effect on their use for model adequacy checking as long as n is not small relative to the number of parameters p.

Methods for scaling residuals

Sometimes it is useful to work with **scaled residuals**. Scaled residuals are helpful in finding observations that are **outliers**, or **extreme values**, i.e., the observations that are separated in some fashion from the rest of the data.

Four popular methods for scaling residuals:

1. Standardized residuals;
2. Studentized residuals;
3. PRESS residuals;
4. *R*-student.

1. Standardized residuals

- **Standardized residuals**:

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \quad i = 1, \ldots, n; \tag{1}$$

The standardized residuals have mean zero and approximately unit variance. Consequently, a large standardized residual ($d_i > 3$, say) potentially indicates an outlier.

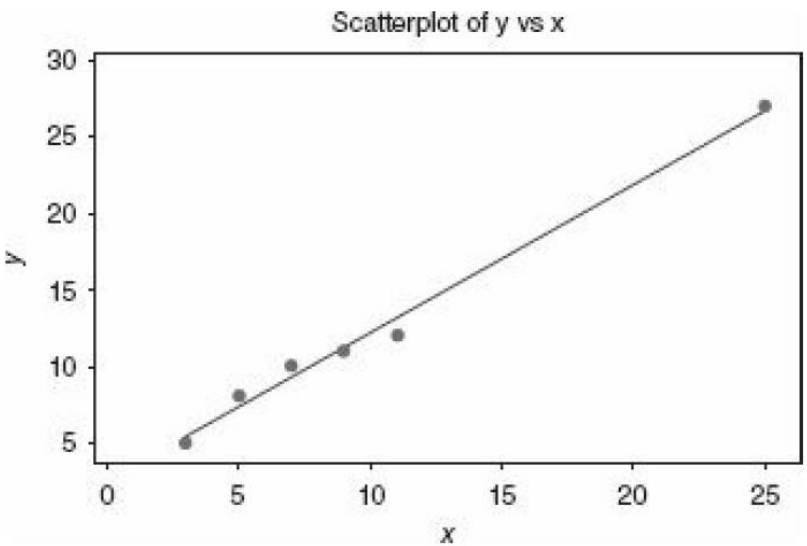
2. Studentized residuals

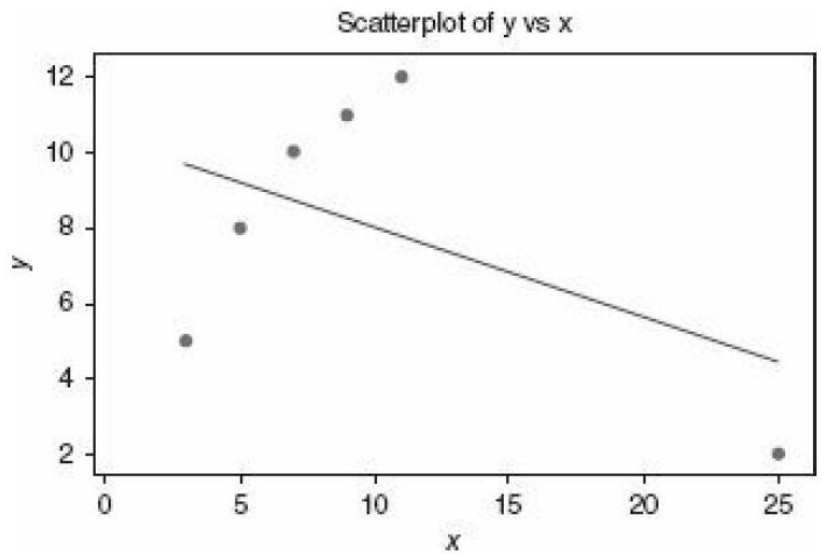
- **Studentized residuals**:

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}},\ i = 1, \ldots, n \tag{2}$$

  - ⋆ To improve the residual scaling by dividing e_i by the exact standard deviation of the ith residual;
  - ⋆ $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\varepsilon$, the residuals are the same linear transformation of the observations $\boldsymbol{y}$ and the error ε;
  - ⋆ $Var(e_i) = \sigma^2(1 - h_{ii})$ and $Cov(e_i, e_j) = -\sigma^2 h_{ij}$;
  - ⋆ $0 \leq h_{ii} \leq 1$, and h_{ii} is called the "**leverage**" and measure the **location** of the ith point in x space;
  - ⋆ If h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response $\hat{y}_i$. On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response $\hat{y}_i$;
  - ⋆ If there is only one regressor, the studentized residuals:
    $r_i = \frac{e_i}{\sqrt{MS_{Res}\left[1-\left(\frac{1}{n}+\frac{(x_i-\hat{x})^2}{S_{xx}}\right)\right]}},\ i = 1, \ldots, n.$

- **Studentized residuals** (cont'd):
  - ⋆ Remark: Generally points near the center of the x space have larger variance (poorer least-squares fit) than residuals at more remote locations.
    - In the left figure, the value of the response is 25, and in the right figure, the value is 2;
    - The left figure is a typical scatter plot for a pure **leverage** point. Such a point is remote in terms of the specific values of the regressors, but the observed value for the response is **consistent** with the prediction based on the other data values.
    - The right figure is a typical scatter plot for an **influential** point. Such a data value is not only remote in terms of the specific values for the regressors, but the observed response is **not consistent** with the values that would be predicted based on only the other data points.





- **Studentized residuals** (cont'd):

Remark:

- The studentized residuals have constant variance $Var(r_i) = 1$ regardless of the location of x_i when the form of the model is correct.
- In many situations the variance of the residuals stabilizes, particularly for large data sets. In these cases there may be little difference between the standardized and studentized residuals.
- Thus, standardized and studentized residuals often convey equivalent information. However, since any point with a large residual **and** a large h_{ii} is potentially highly influential on the least-squares fit, examination of the studentized residuals is generally recommended.
- It is customary to use MS_{Res} as an estimate of σ^2 in computing r_i. This is referred to as **internal scaling** of the residual because MS_{Res} is an internally generated estimate of σ^2 obtained from fitting the model to all n observation.

3. PRESS residuals

- **PRESS residuals** (also called **prediction errors** or **deleted residuals**):

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}}, \ i = 1, \ldots, n, \tag{3}$$

where $\hat{y}_{(i)}$ is the fitted value of ith response based on all observation except the ith one.

  - ⋆ The logic is that if the ith observation y_i is really unusual, the regression model based on all observations may be overly influenced by this observation. This could produce a fitted value $\hat{y}_i$ that is very similar to the observed value y_i, and consequently, the ordinary residual e_i will be small.
  - ⋆ Residuals associated with points for which h_{ii} is large will have large PRESS residuals. These points will generally be **high influence** points.
  - ⋆ Generally, a large difference between the ordinary residual and the PRESS residual will indicate a point where the model fits the data well, but a model built without that point **predicts** poorly.
  - ⋆ The variance of the ith PRESS residual is $Var(e_{(i)}) = \frac{\sigma^2}{1-h_{ii}}$, so that a standardized PRESS residual is

$$\frac{e_{(i)}}{\sqrt{Var[e_{(i)}]}} = \frac{e_i}{\sqrt{\sigma^2(1 - h_{ii})}}$$

If we use MS_{Res} to estimate σ^2, is just the studentized residual.

4. *R*-student

- ***R*-student** (also called **externally studentized residual**):

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}\left(1 - h_{ii}\right)}}, \quad i = 1, \ldots, n, \tag{4}$$

where $S^2_{(i)} = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-p-1}$.

  - ⋆ This approach would be to use an estimate of σ^2 based on a data set with the ith observation removed.
  - ⋆ In many situations t_i will differ little from the studentized residual r_i. However, if the ith observation is influential, then can differ significantly from MS_{Res}, and thus the R-student statistic will be more sensitive to this point.
  - ⋆ $t_i \sim t_{n-p-1}$.
  - ⋆ In general, a **diagnostic view** as opposed to a strict statistical hypothesis testing view is best.

Example 1: the delivery time data

- A soft drink bottler is analyzing the vending machine service routes in his distribution system, and he is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet.
- y: the delivery time; x_1: the No. of cases of product stocked; x_2: the distance walked by the route driver.
- The least-squares fit is $\hat{y} = 2.3412 + 1.6159x_1 + 0.0144x_2$.

Observation No., i	Delivery Time, y_i (psi)	No. of Cases, x_1	Distance, x_2
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

Table 1: Scaled residuals for Example 1

Observation	$e_i = y_i - \hat{y}_i$	$d_i = e_i \Big/ \sqrt{MS_{Res}}$	$r_i = e_i \Big/ \sqrt{MS_{Res}(1-h_{ii})}$	h_{ii}	$e_{(i)} = e_i \Big/ (1-h_{ii})$	$t_i = e_i \Big/ \sqrt{S^2_{(i)}(1-h_{ii})}$	$[e_i \Big/ (1-h_{ii})]^2$
Number, i	(1) the ordinary residuals	(2) standardized residuals	(3) studentized residuals	(4)	(5) residuals residuals	(6) R-student	(7)
1	-5.0281	-1.5426	-1.6277	0.10180	-5.5980	-1.6956	31.3373
2	1.1464	0.3517	0.349	0.07070	1.2336	0.3575	1.5218
3	-0.0498	-0.0153	-0.0161	0.09874	-0.0557	-0.0157	0.0031
4	4.9244	1.5108	1.5798	0.05838	5.2297	1.6392	27.3499
5	-0.4444	0.1363	0.1418	0.07501	-0.4804	-0.1386	0.2308
6	-0.2896	-0.0888	-0.0908	0.04287	-0.3025	-0.0887	0.0915
7	0.8446	0.2501	0.2704	0.08180	0.0918	0.2646	0.8461
8	1.1566	0.3548	0.3667	0.06373	1.2353	0.3594	1.5260
9	**7.4197**	**2.2763**	**3.2138**	**0.49829**	**14.7888**	**4.3108**	**218.7093**
10	2.3764	0.7291	0.8133	0.19630	2.9568	0.8068	8.728
11	2.2375	0.6865	0.7181	0.08613	2.4484	0.7099	5.9946
12	-0.5930	-0.1819	-0.1932	0.11366	-0.6690	-0.1890	0.4476
13	1.0270	0.3151	0.3252	0.06113	1.0938	0.3185	1.1965
14	1.0675	0.3275	0.3411	0.07824	1.1581	0.3342	1.3412
15	0.6712	0.2059	0.2103	0.04111	0.7000	0.2057	0.4900
16	-0.6629	-0.2034	-0.2227	0.16594	-0.7948	-0.2178	0.6317
17	0.4364	0.1339	0.1381	0.05943	0.4640	0.1349	0.2153
18	3.4486	1.0580	1.1130	0.09626	3.8159	1.1193	14.5612
19	1.7932	0.5502	0.5787	0.09645	1.9846	0.5698	3.9387
20	-5.7880	-1.7758	-1.8736	0.10169	-0.6432	-1.9967	41.5150
21	-2.6142	-0.8020	-0.8779	0.16528	-3.1318	-0.8731	9.8084
22	-3.6865	-1.1310	-1.4500	0.39158	**-6.0591**	-1.4896	36.7131
23	-4.6076	-1.4136	-1.4437	0.04126	-4.8059	-1.4825	23.0966
24	-4.5728	-1.4029	-1.4961	0.12061	-5.2000	-1.5422	27.0397
25	-0.2126	-0.0652	-0.0675	0.06664	-0.2278	-0.0660	0.0519
							PRESS=457.4000

Example 1

- Examining column (1) we note that one residuals $e_9 = 7.4197$, seems suspiciously large.
- Column (2) shows that the standardized residual $d_9 = 2,2763$. All other standardized residuals are inside the ± 2 limits.
- Column (3) shows the studentized residual at point 9 is $r_9 = 3.2138$, which is substantially larger than the standardized residuals.
- Column (4) shows the diagonal elements of the hat matrix, which are used extensively in computing scaled residuals.
- Column (5) contains the PRESS residuals. The PRESS residuals for point 9 and 22 are substantially larger than the corresponding ordinary residuals, *indicating that these are likely to be points where the model fits reasonably well but does not provide good predictions of fresh data.*
- Column (6) displays the value of R-student Only t_9 is unusually large. Note that t_9 is larger than the corresponding studentized residuals r_9, *indicating that when run* 9 *is set aside,* $S^2_{(9)}$ *is smaller than* MS_{Res}*, so clearly this run is influential.*

  Note that

$$S^2_{(9)} = \frac{(n-p)MS_{Res} - e_9^2/(1-h_{99})}{n-p-1} = \frac{(22)(10.6239) - (7.4197)^2/(1-0.49829)}{21}$$

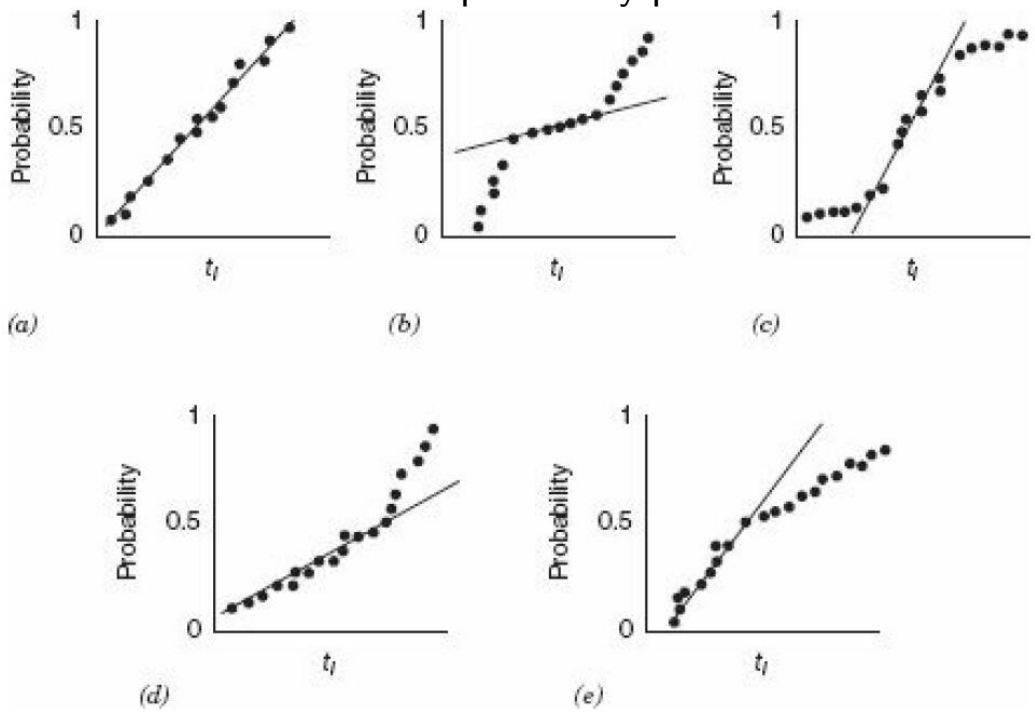
$$= 5.9046 < 10.6239 = MS_{Res}.$$

Residual plots

Graphical analysis of residuals is a very effective way to investigate the adequacy of the fit of a regression model and to check the underlying assumptions.

- **Normal probability plot**
  - ⋆ Reason:
    - Gross nonnormality is potentially more serious as the t or F statistics and confidence and prediction intervals depend on the normality assumption.
    - Furthermore, if the errors come from a distribution with thicker or heavier tails than the normal, the least-squares fit may be sensitive to a small subset of the data. Heavy-tailed error distributions often generate outlier that "pull" the least-squares fit too much in their direction.

Normal probability plot



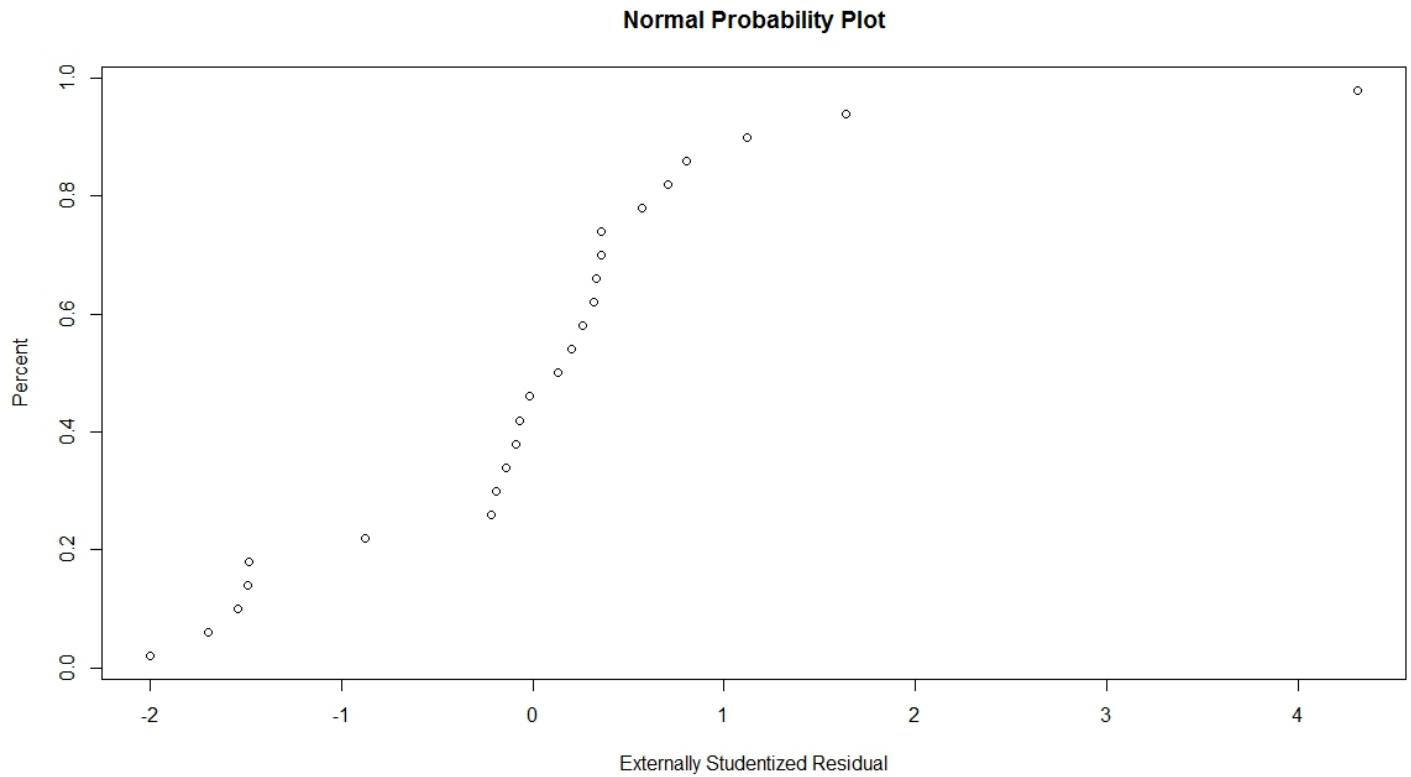
- The normal probability plots are constructed by plotting the "expected normal value" $\phi^{-1}[(i-\frac{1}{2})/n]$ against the ranked residual $t_{[i]}$, where ϕ denotes the standard normal cumulative distribution.
- Panel (a) displays an "idealized" **normal probability plot**. Notice that the points lie approximately along a straight line;
- Panel (b) shows a sharp upward and downward curve at both extremes, indicating that the tails of this distribution are too light for it to be considered normal (**light tailed**).
- Conversely, panel (c) shows flattening at the extremes, which is a pattern typical of samples from a distribution with heavier tails than the normal (**heavy tailed**).
- Panels (d) and (e) exhibit patterns associated with **positive** and **negative skew**, respectively.

Remark:

- Since samples taken from a normal distribution will not plot exactly as a straight line, some experience is required to interpret normal probability plots.
- Study of these plots is helpful in acquiring a feel for how much deviation from the straight line is acceptable.
- A common defect that shows up on the normal probability plot is the occurrence of one or two large residuals and this is an indication that the corresponding observations are **outliers**.

Example 2: the delivery time data

```
> #### Example 4.2
> setwd("C:/Users/user/Desktop/AMA514_2016/AMA514_2016_new/Data")
> E3.3<-read.table("data_example_3.1.txt", header=T)
> ## Simplify notation
> y<-E3.3$Time; x1<-E3.3$Cases; x2<-E3.3$Distance
> ## Normal probability plot
> res<-lm(y~x1+x2)
> r=rstudent(res)
> qqplot(sort(r), ppoints(res$fit), xlab='Externally Studentized Residual', ylab='Percent', main="Normal Probability Plot")
```



- The residuals do not lie exactly along a straight line, indicating that there may be some problems with the normality assumption, or that there may be one or more outliers in the data.
- From Example 1, we know that the studentized residual for observation 9 is moderately large ($r_9 = 3.2138$), as is the R-student residual ($t_9 = 4.3108$). However, there is no indication of a severe problem in the delivery time data.

Plot of residuals against the fitted values $\hat{y}_i$

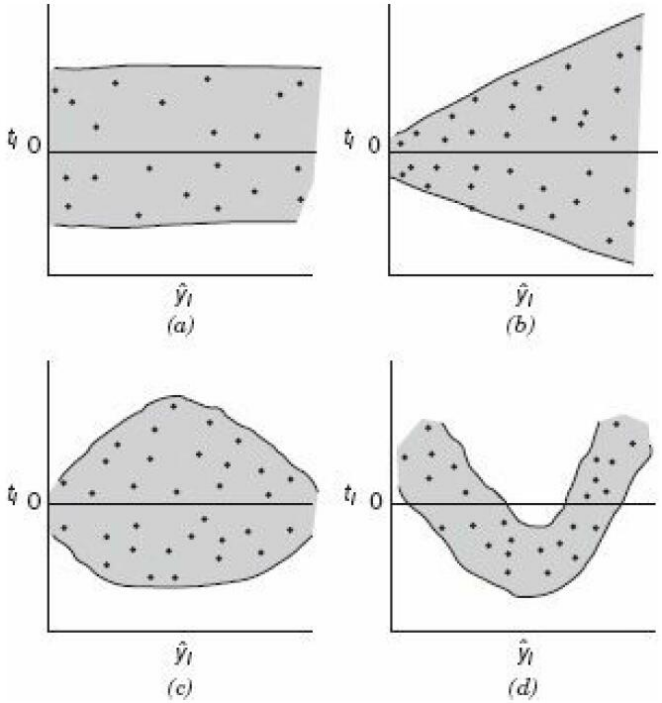


Figure: Patterns for residuals plots: (a) satisfactory; (b) funnel; (c) double bow; (d) nonlinear.

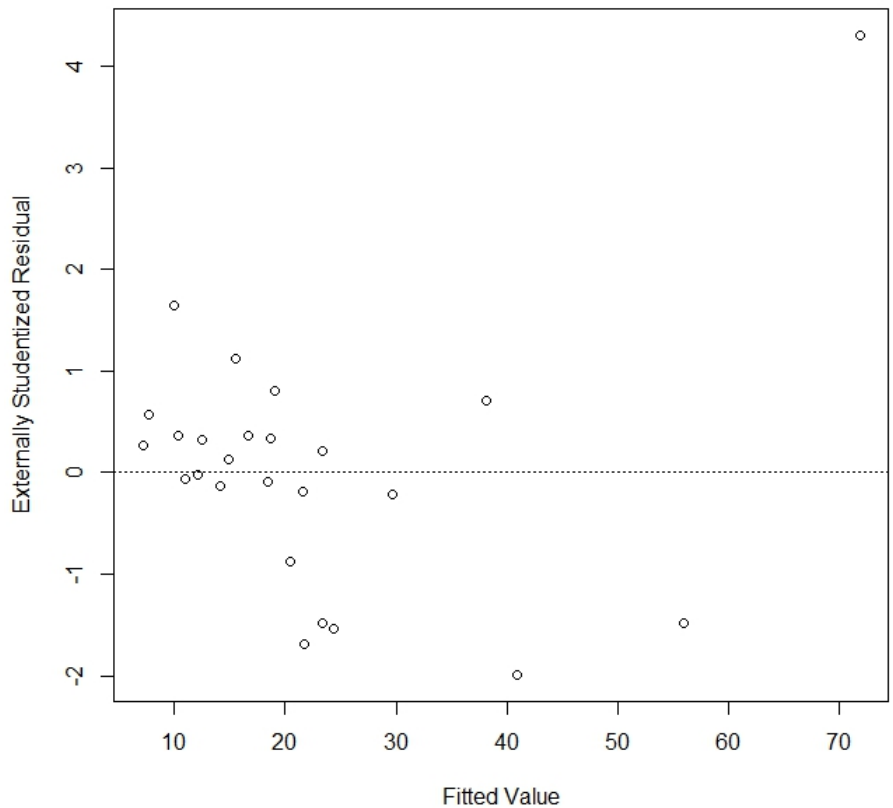
- If this plot resembles Figure (a), which indicates that the residuals can be contained in a **horizontal band**, then there are no obvious model defects.
- Plots of t_i versus $\hat{y}_i$ that resemble any of the patterns in panels $(b) - (d)$ are symptomatic of model deficiencies.
- The patterns in panels (b) and (c) indicate that the variance of the errors is not constant.
- The outward-opening **funnel** pattern in panel (b) implies that the variance is an increasing function of y.
- The **double-bow** pattern in panel (c) often occurs when y is a proportion between 0 and 1. The variance of a binomial proportion near 0.5 is greater than one near 0 or 1.
- A **curved** plot such as in panel (d) indicates nonlinearity. This could mean that other regressor variables are needed in the model.

Remark:

- A plot of the residuals against $\hat{y}_i$ may also reveal one or more unusually large residuals. These points are, of course, potential outliers.
- Large residuals that occur at the extreme $\hat{y}_i$ values could also indicate that either the variance is not constant or the true relationship between y and x is not linear.
- These possibilities should be investigated before the points are considered outliers.

```
> ## Plot of externally studentized residuals vs. predited value
> plot(res$fit, r, xlab="Fitted Value", ylab="Externally Studentized Residual")
> abline(h=0, lty=3, lwd=1)
```

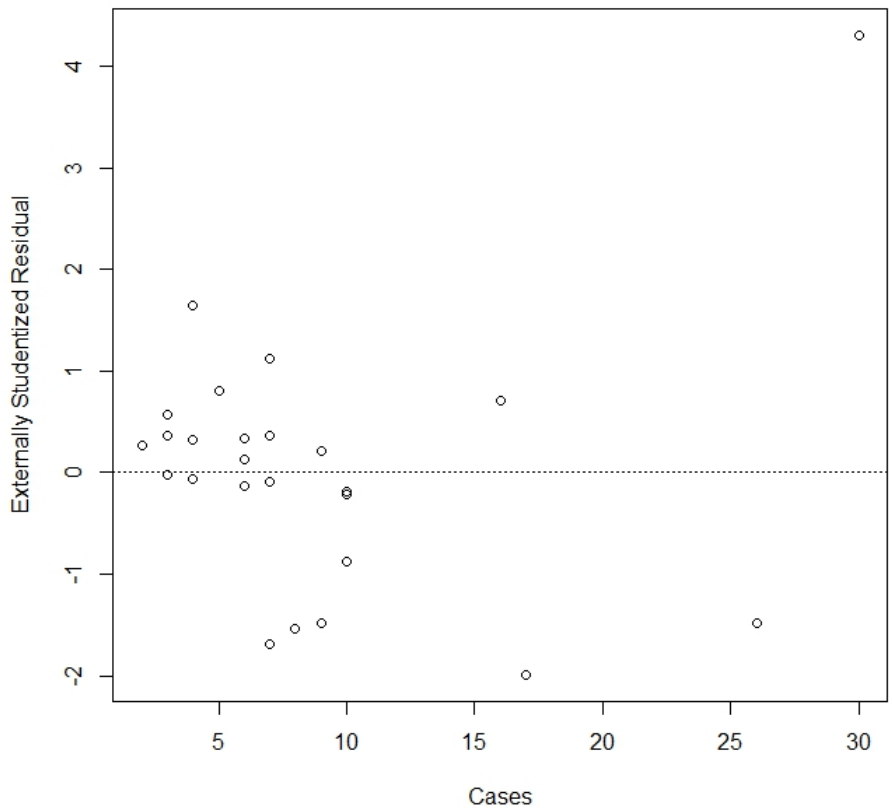
Versus Fits (response is time)



- The plot does not exhibit any strong unusual pattern, although the large residual t_9 shows up clearly.
- There does seem to be a slight tendency for the model to underpredict short delivery times and overpredict long delivery times.

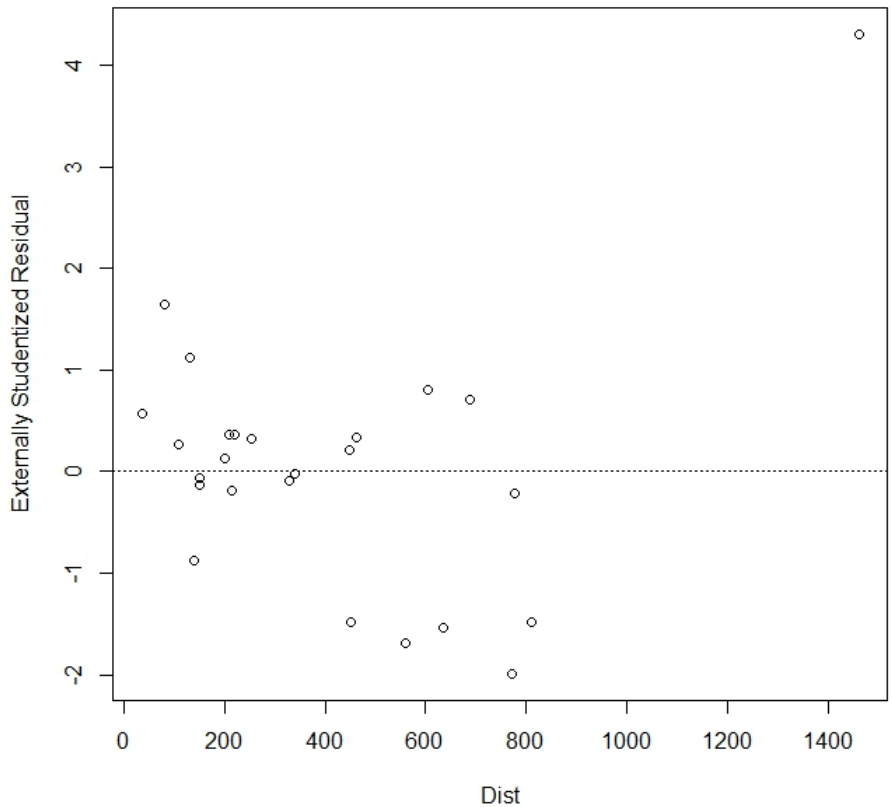
- **Plot of residuals against the regressor**

```
> ## Plot of externally studentized residuals vs. predited value
> plot(x1, r, xlab="Cases", ylab="Externally Studentized Residual")
> abline(h=0, lty=3, lwd=1)
```



⋆ This figure plots residuals versus cases

```
> ## Plot of externally studentized residuals vs. predited value
> plot(x2, r, xlab="Dist", ylab="Externally Studentized Residual")
> abline(h=0, lty=3, lwd=1)
```



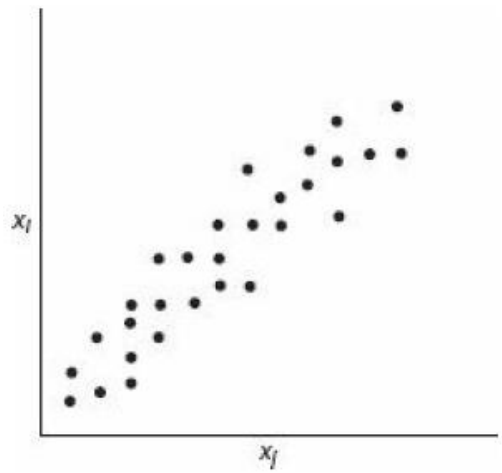
- This figure plots residuals versus distance.

Remark:

- Neither of these plots reveals any clear indication of a problem with either misspecification of the regressor (implying the need for either a transformation on the regressor or higher order terms in cases and/or distance) or inequality of variance, although the moderately large residual associated with point 9 is apparent on both plots.
- It is also helpful to plot residuals against regressor variables that are not currently in the model but which could potentially be included. Any structure in the plot of residuals versus an omitted variable indicates that incorporation of that variable could improve the model.
- Plotting residuals versus a regressor is not always the most effective way to reveal whether a curvature effect (or a transformation) is required for that variable in the model.

Other residual plotting and analysis methods

- In addition to the basic residuals plots discussed previously, there are several others that are occasionally useful.
  - ⋆ a scatterplot between x_i against regressor x_j, which may be useful in studying the relationship between regressor variables and the disposition of the data in x space.



- This display indicates that x_i and x_j are highly positively correlated.
- Plots of x_i versus x_j may also be useful in discovering points that are remote from the rest of the data and that potentially influence key model properties.

PRESS Statistic

- PRESS residuals: $e_{(i)} = y_i - \hat{y}_{(i)}$, where $\hat{y}_{(i)}$ is the predicted value of the ith observed response based on a model fit to the remaining $n-1$ sample points.
- We noted that large PRESS residuals are potentially useful in identifying observations where the model does not fit the data well or observations for which the model is likely to provide poor future predictions.
- The PRESS statistic:

$$\begin{aligned} \text{PRESS} &= \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2 \\ &= \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 \end{aligned}$$

- *PRESS is generally regarded as a measure of how well a regression model will perform in predicting new data.* A model with a **small** value of PRESS is desired.
- R^2 for prediction based on PRESS:

$$R^2_{\text{prediction}} = 1 - \frac{\text{PRESS}}{SS_T}$$

  This statistic gives some indication of the predictive capability of the regression model.
- Using PRESS to compare models: One very important use of the PRESS statistic is in comparing regression models. Generally, a model with a **small** value of PRESS is preferable to one where PRESS is large.

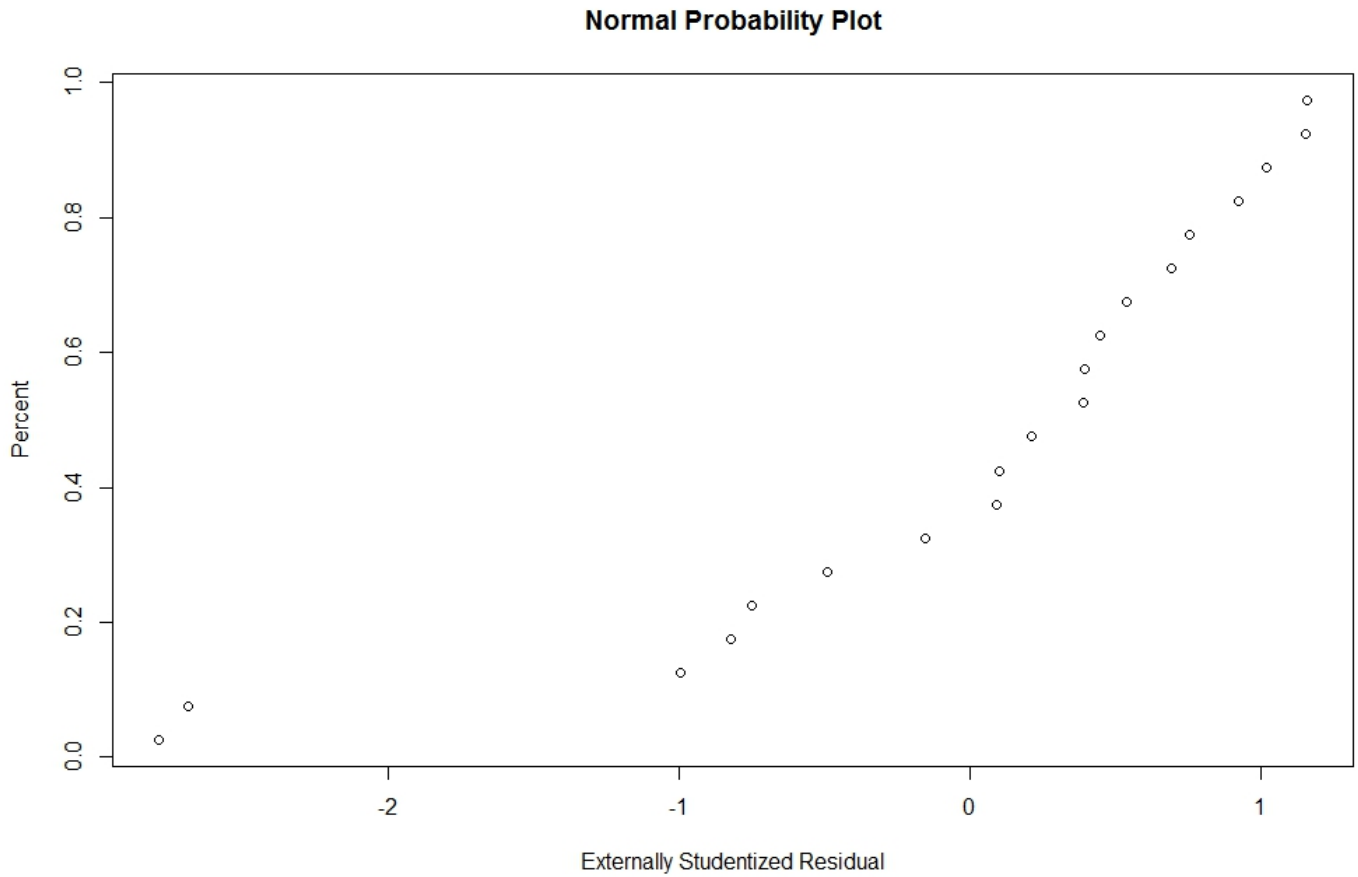
Detection and treatment of outliers

- An outlier is an extreme observation
  - ⋆ One that is considerably different from the majority of the data;
  - ⋆ Residuals that are considerably larger in absolute value than the others, say three or four standard deviations from the mean, indicate potential y space outliers.
  - ⋆ Outliers are data points that are not typical of the rest of the data. Depending on their location in x space, outliers can have moderate to severe effects on the regression model;
  - ⋆ Examining **scaled residuals**, such as the studentized and R-student residuals, is an excellent way to identify potential outliers.
- Outliers should be carefully investigated to see if a reason for their unusual behavior can be found.
  - ⋆ Sometimes outliers are "bad" values, occurring as a result of unusual but explainable events;
  - ⋆ Clearly discarding bad values is desirable because least squares pulls the fitted equation toward the outlier as it minimizes the residual sum of squares;
  - ⋆ However, we emphasize that there should be strong nonstatistical evidence that the outlier is a bad value before it is discarded

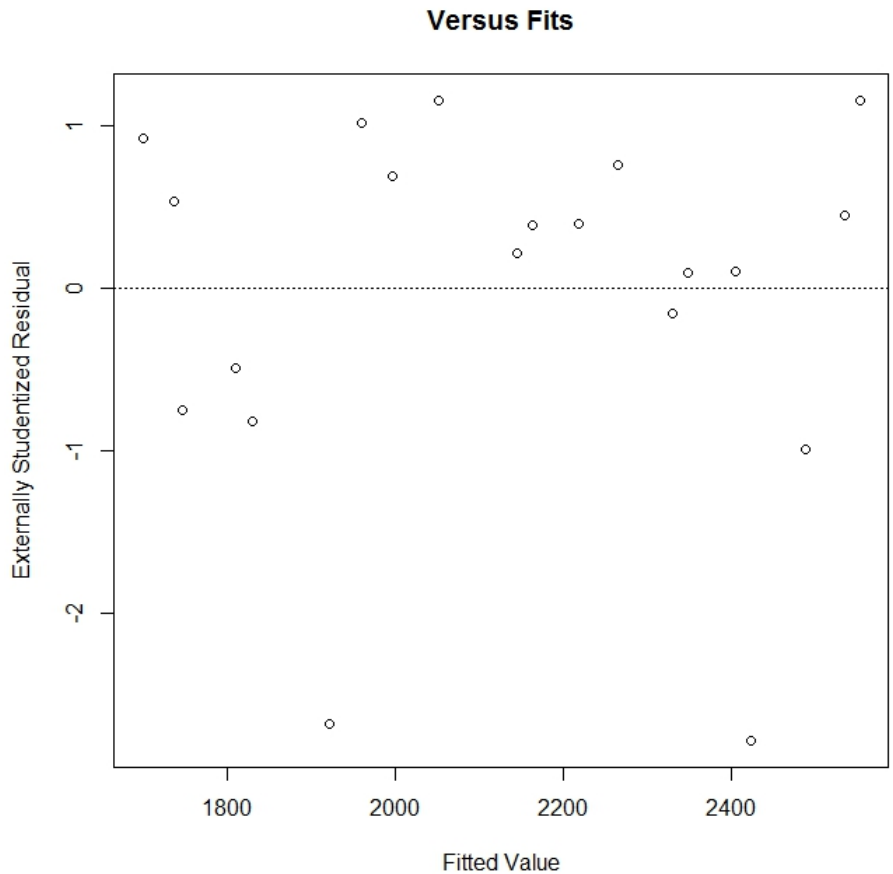
- Sometimes we find that the outlier is an unusual but perfectly plausible observation.
  - ⋆ Deleting these points to improve the fit of the equation can be dangerous, as it can give the user a false sense of precision in estimation or prediction;
  - ⋆ Outliers may also point out inadequacies in the model, such as failure to fit the data well in a certain region of x space.
  - ⋆ If the outlier is a point of particularly desirable response, knowledge of the regressor values when that response was observed may be extremely valuable.
  - ⋆ Identification and follow-up analyses of outliers often result in process improvement or new knowledge concerning factors whose effect on the response was previously unknown.
- The effect of outliers on the regression model may be easily checked by dropping these points and refitting the regression equation.
  - ⋆ We may find that the values of the regression coefficients or the summary statistics such as the t or F statistic, R^2, and the residual mean square may be very sensitive to the outliers.
  - ⋆ Situations in which a relatively small percentage of the data has a significant impact on the model may not be acceptable to the user of the regression equation.
  - ⋆ Generally we are happier about assuming that a regression equation is valid if it is not overly sensitive to a few observations.
  - ⋆ We would like the regression relationship to be embedded in all of the observations and not merely an artifice of a few points.

Example: the rocket propellant data

```
> #### Example 4.7
> setwd("C:/Users/user/Desktop/AMA514_2016/AMA514_2016_new/Data")
> E4.7<-read.table("data_example_2.1.txt", header=T)
> ## Simplify notation
> y<-E4.7$y; x<-E4.7$x
> ## Normal probability plot
> res<-lm(y~x)
> r=rstudent(res)
> qqplot(sort(r), ppoints(res$fit), xlab='Externally Studentized Residual', ylab='Percent', main="Normal Probability Plot")
```



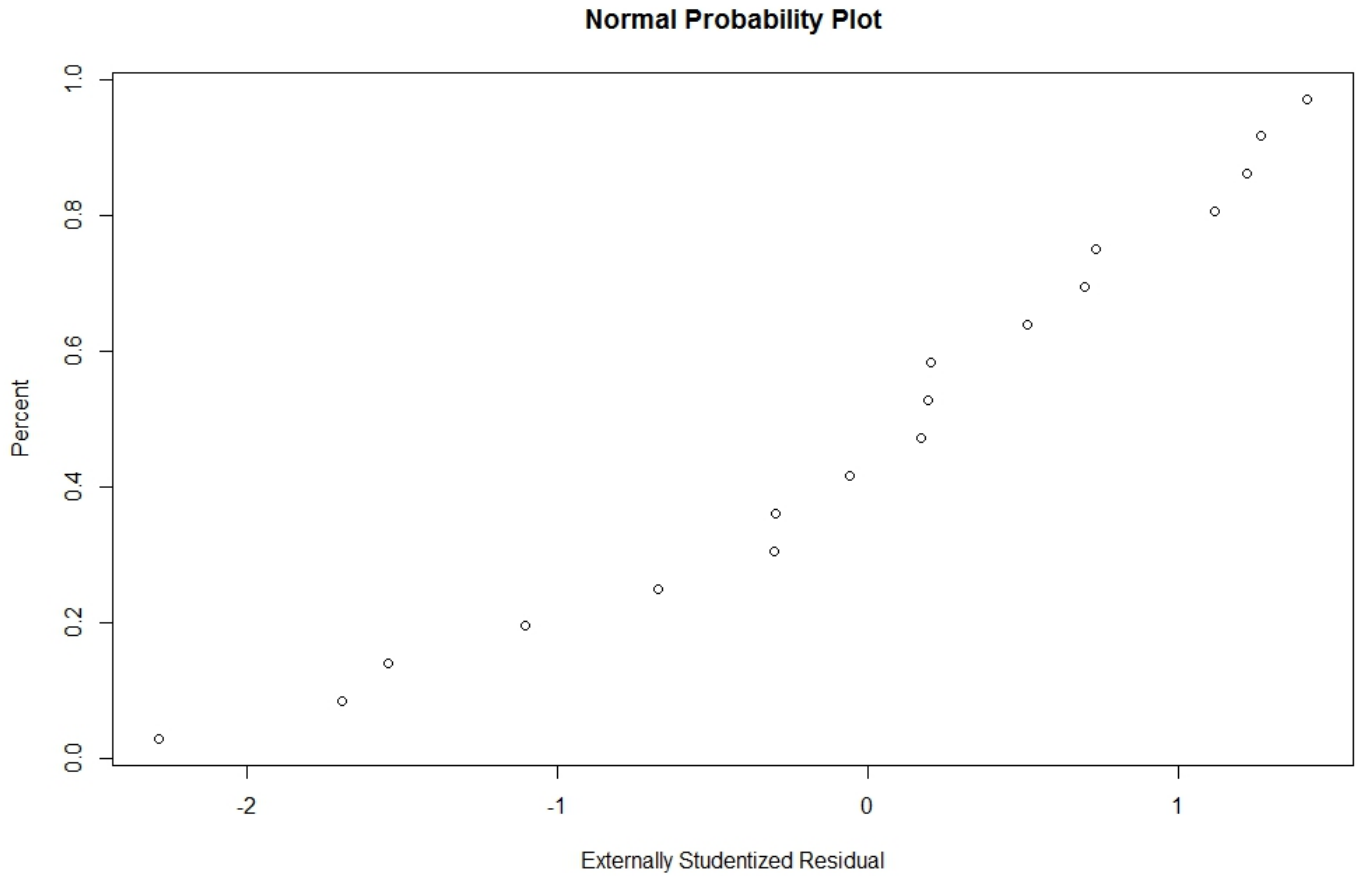
```
> ## Plot of externally studentized residuals vs. predited value
> plot(res$fit, r, xlab="Fitted Value", ylab="Externally Studentized Residual", main="Versus Fits")
> abline(h=0, lty=3, lwd=1)
```



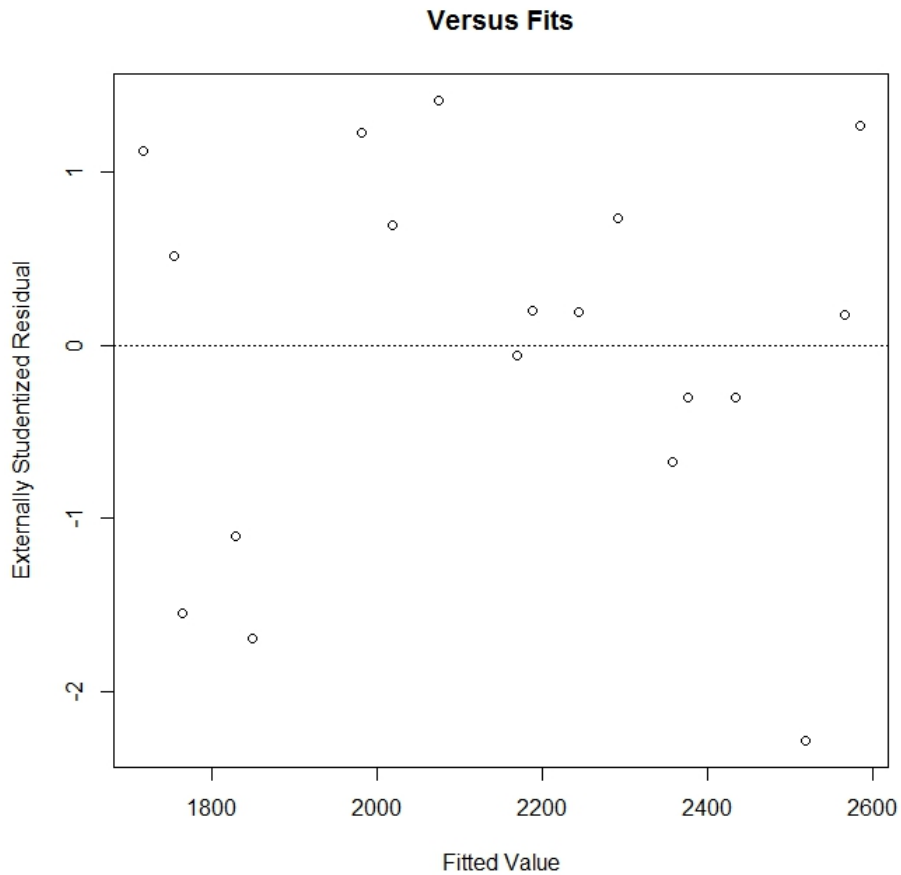
```
> summary(res)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-215.98  -50.68   28.74   66.61  106.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16 ***
x            -37.154      2.889  -12.86 1.64e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10

> anova(res)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1527483 1527483  165.38 1.643e-10 ***
Residuals 18  166255    9236
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Delete points 5 and 6.

```
> E4.71<-E4.7[-c(5,6),]
> ## Simplify notation
> y1<-E4.71$y; x1<-E4.71$x
> ## Normal probability plot
> res1<-lm(y1~x1)
> r1=rstudent(res1)
> qqplot(sort(r1), ppoints(res1$fit), xlab='Externally Studentized Residual', ylab='Percent', main="Normal Probability Plot")
```



```
> ## Plot of externally studentized residuals vs. predited value
> plot(res1$fit, r1, xlab="Fitted Value", ylab="Externally Studentized Residual", main="Versus Fits")
> abline(h=0, lty=3, lwd=1)
```



```
> summary(res1)

Call:
lm(formula = y1 ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-118.07  -35.67   11.31   44.75   83.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2658.973     30.533   87.08  < 2e-16 ***
x1           -37.694      1.979  -19.05 2.02e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.97 on 16 degrees of freedom
Multiple R-squared:  0.9578,    Adjusted R-squared:  0.9551
F-statistic: 362.9 on 1 and 16 DF,  p-value: 2.023e-12

> anova(res1)
Analysis of Variance Table

Response: y1
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 1438842 1438842  362.92 2.023e-12 ***
Residuals 16   63434    3965
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remark:

- Since the estimates of the parameters have not changed dramatically, we conclude that points 5 and 6 are not overly influential. They lie somewhat off the line passing through the other 18 points, but they do not control the slope and intercept.
- However, these two residuals make up approximately 56% of the residual sum of squares.
- Thus, if these points are truly bad values and should be deleted, the precision of the parameter estimates would be improved and the widths of confidence and prediction intervals could be substantially decreased.