Section Outline

1. Non-Constant Variation
2. Outliers

Non-Constant Variation

Non-constant variation across the levels of the predictor variables violates one of the usual assumptions of least squares regression.

However non-constant variation is often found in data used for building regression models as an inherent part of the measurement or other type of process.

Because non-constant variation is frequently encountered, techniques to improve the validity of the standard assumptions and techniques based on alternative assumptions have been developed to lessen the impact of this problem.

Effect of Non-Constant Standard Deviation on Function Fitting

If all of the other assumptions for the analysis are met, except for constant standard deviation across all combinations of the predictors, the usual parameter estimates will still be correct on average, or unbiased.

However, the deviation of the fitted function from the true function will be larger, on average, than that from other techniques, unlike the case of constant standard deviation.

Effect of Non-Constant Standard Deviation on Function Fitting

The least squares estimates are found by minimizing the sum of the squared differences of the predicted values and the observed y's, Q, to find the $\hat{\beta}$'s.

$$Q = \sum_{i=1}^{n} [y_i - f(x_{1i}, x_{2i}, \ldots, x_{ki}; \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)]^2$$

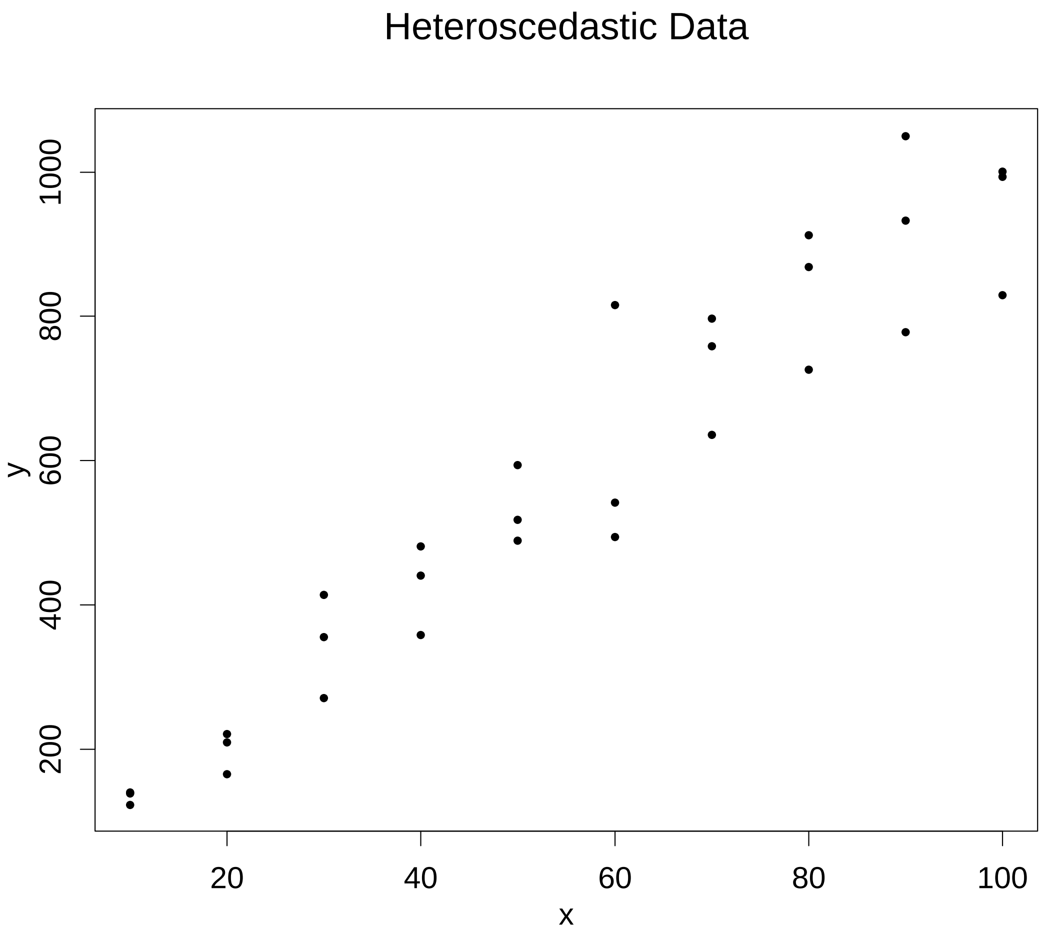
All deviations count the same using the LSS fitting criterion, even if the variation differs across the data.

However, where the variation is larger, the deviations of the data from their corresponding means will also be larger, indicating that the regression function should not be constrained to be as near that data as it is to data that varies less.

Effect of Non-Constant Standard Deviation on Prediction

In addition to increasing the variance of estimating the regression function, non-constant standard deviation impacts the uncertainty of predictions, if unaccounted for.

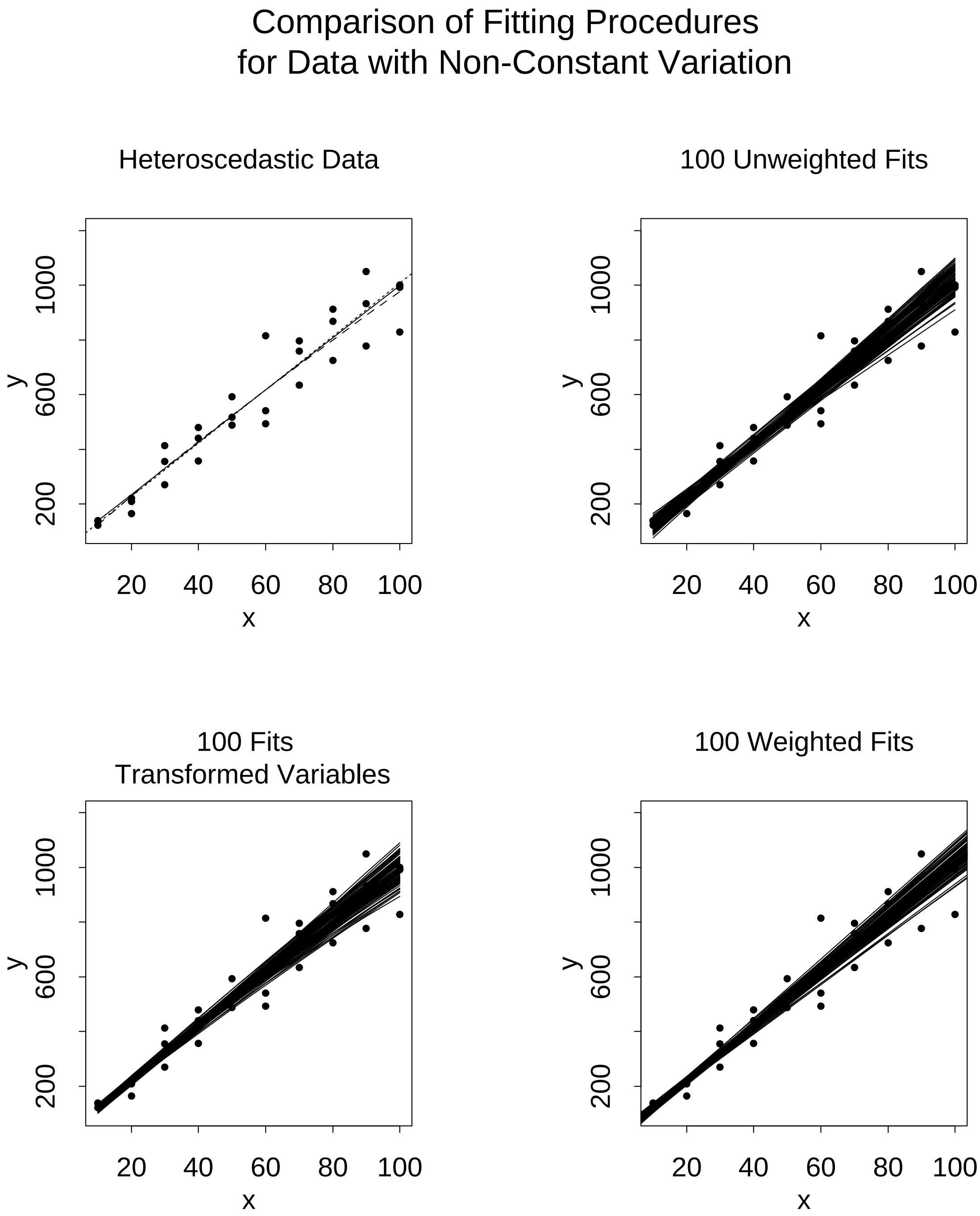
Predictions made where variation is relatively low, will tend to have larger stated uncertainties than necessary. Predictions made where uncertainty is relatively large will have smaller stated uncertainties than necessary.


Heteroscedastic Data
y
x
200
400
600
800
1000
20
40
60
80
100

Improved Fitting Methods

There are basically two approaches to getting improved parameter estimates for data with non-constant standard deviations:

1. transformation of the data so it meets the standard assumptions, and
2. use of weights in parameter estimation to account for the unequal standard deviations.


Comparison of Fitting Procedures
for Data with Non-Constant Variation
Heteroscedastic Data
100 Unweighted Fits
100 Fits
Transformed Variables
100 Weighted Fits
y
x
200
600
1000
20
40
60
80
100

Transformations

The basic steps for using transformations to handle data with unequal subpopulation standard deviations are:

1. Transform the response variable to equalize the variation across the levels of the predictor variables
2. Transform the predictor variables if necessary to attain or restore a simple functional form for the regression function.
3. Fit and validate the model in the transformed variables.
4. Transform the function back into the original units, if necessary, using the inverse of the transformation originally applied to the response variable.

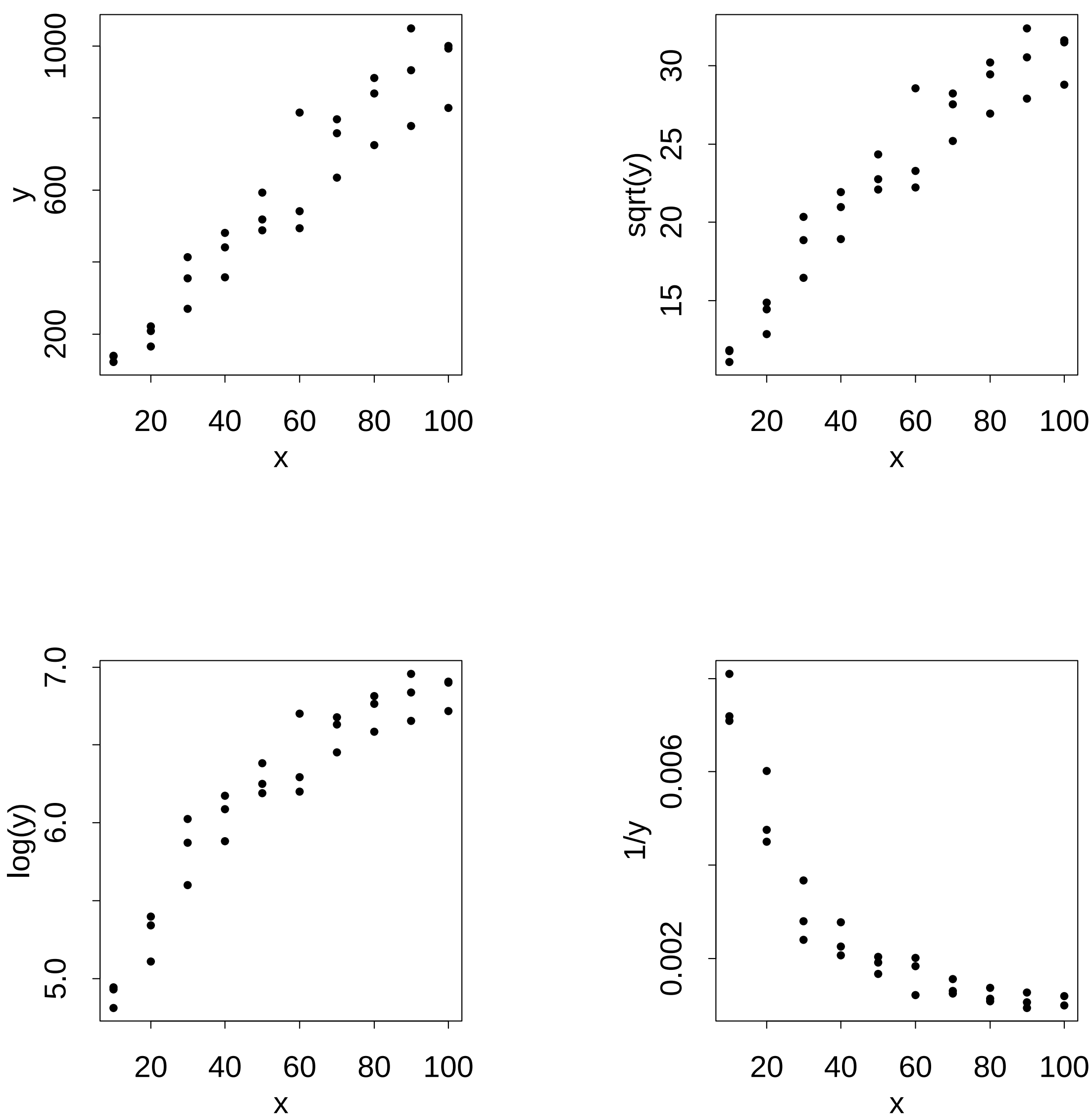
Transformations

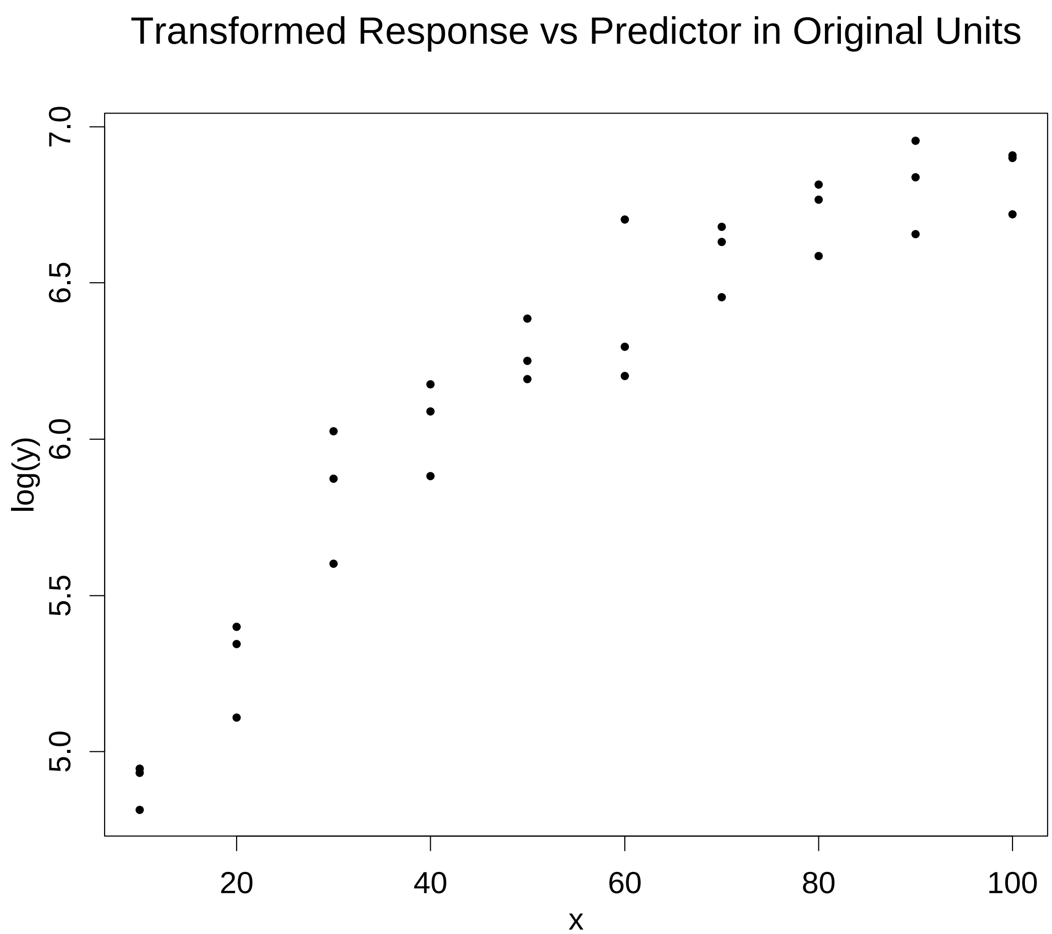
Appropriate transformations to stabilize the variability may be suggested by scientific knowledge or selected using the data.

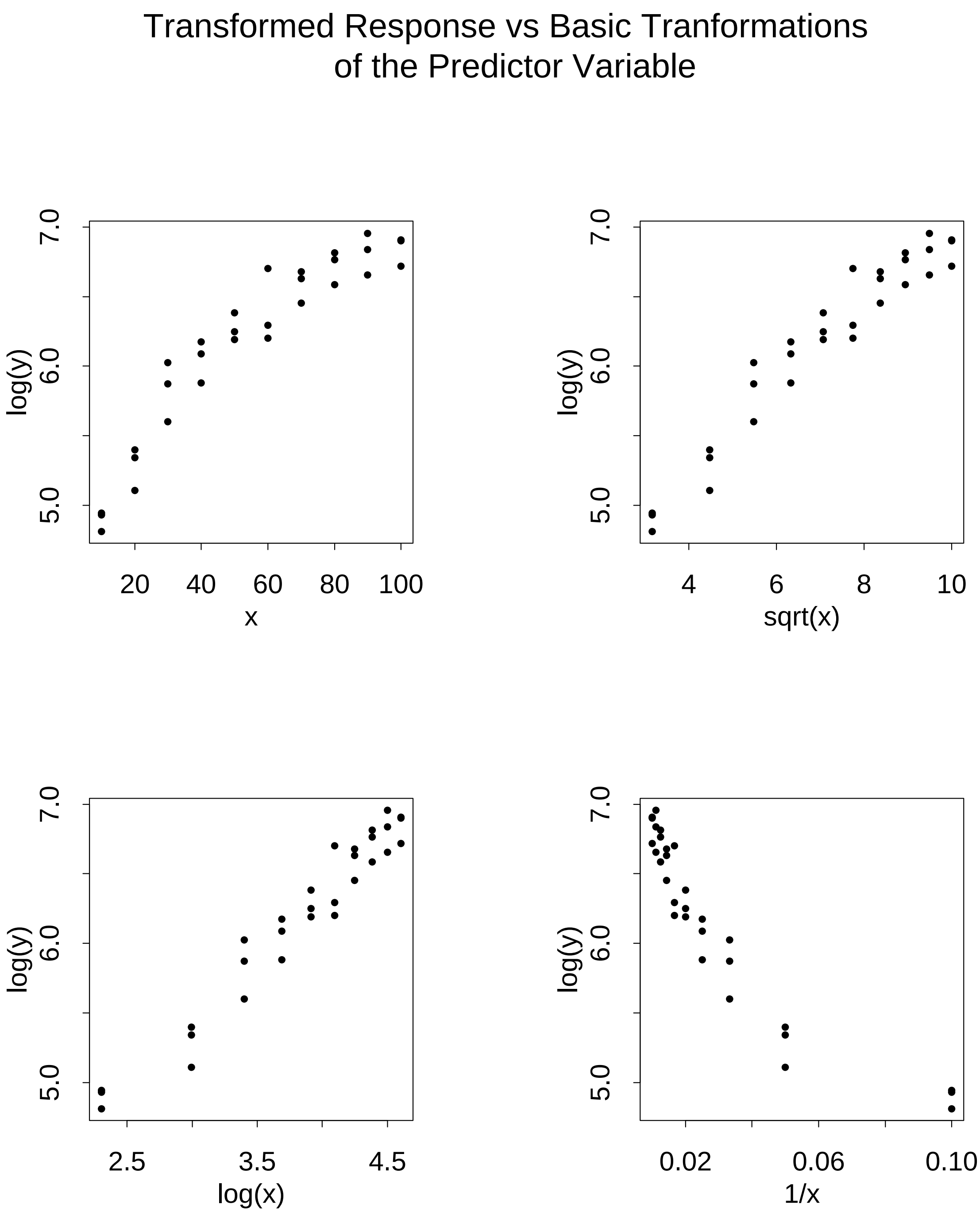
Three transformations that are often effective is equalizing the standard deviations across the values of the predictor are:

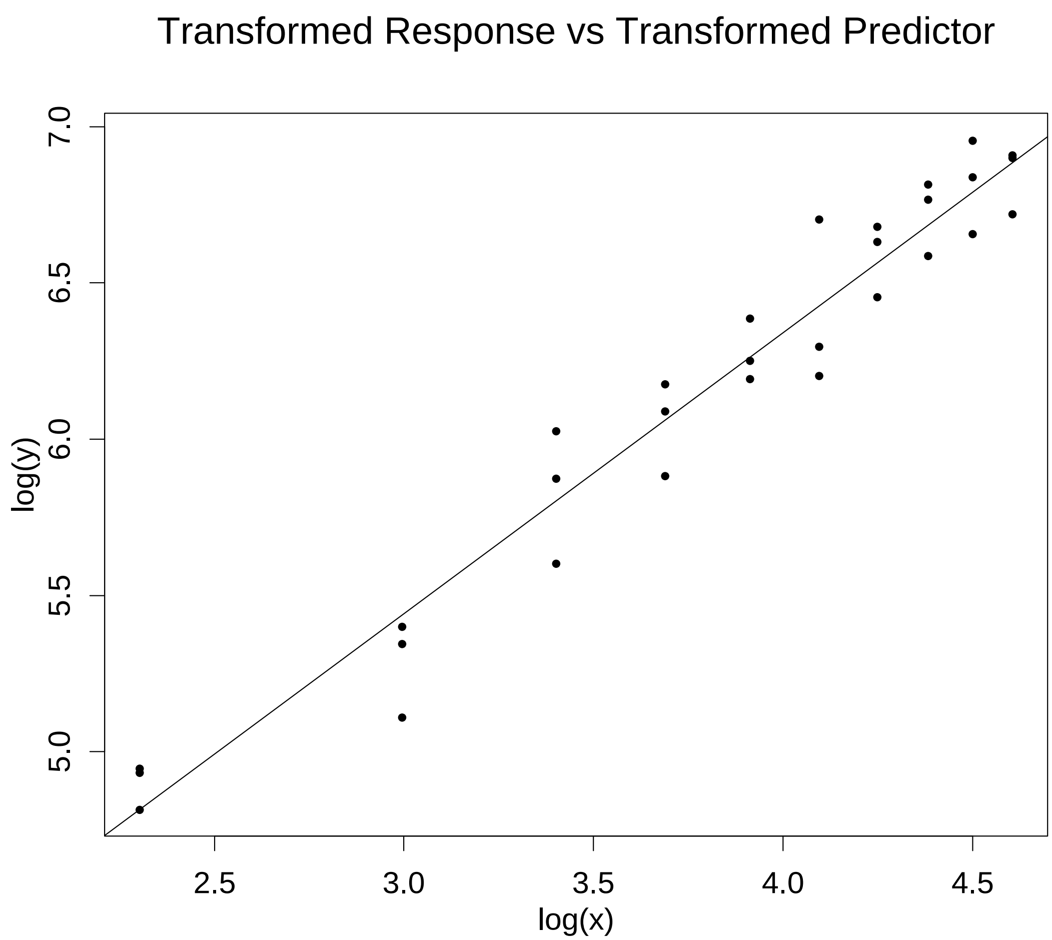
1. $\sqrt{y}$
2. $\log(y)$, and
3. $1/y$.

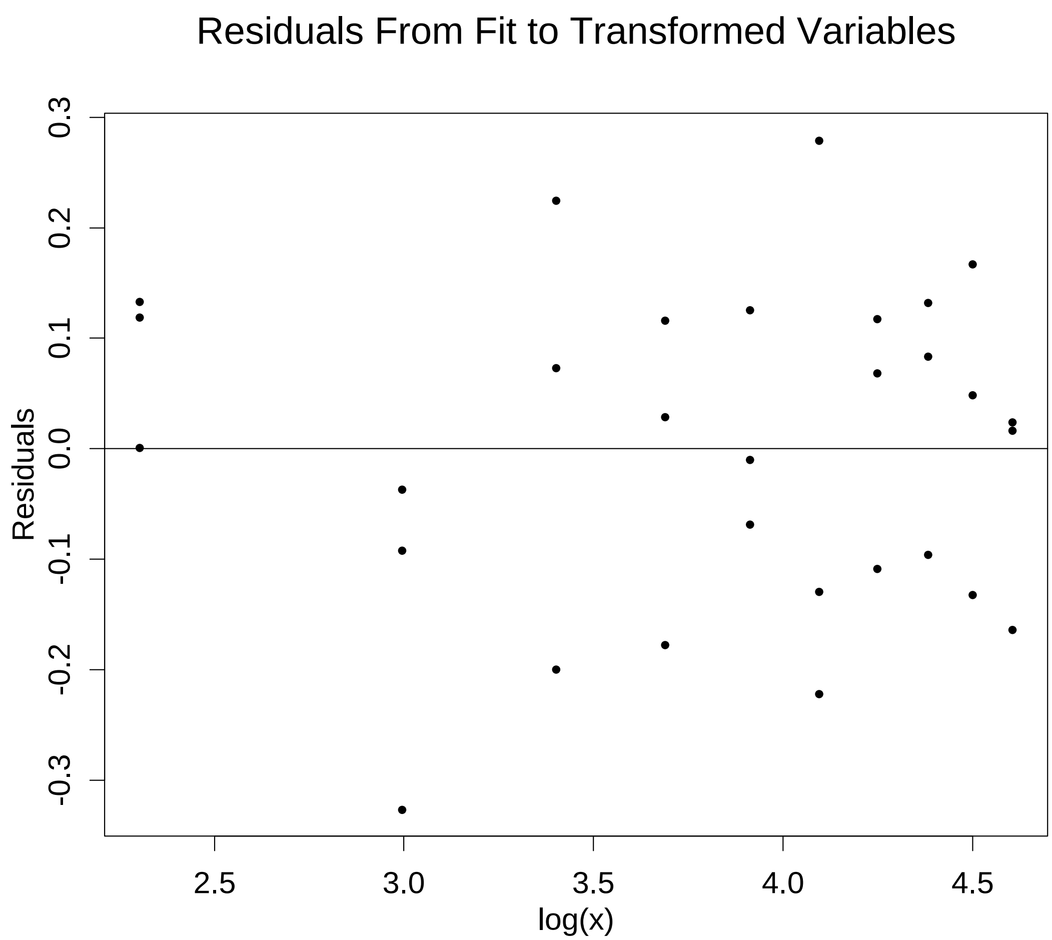
Data in the Original Units and 3 Basic Tranformations
of the Response Variable

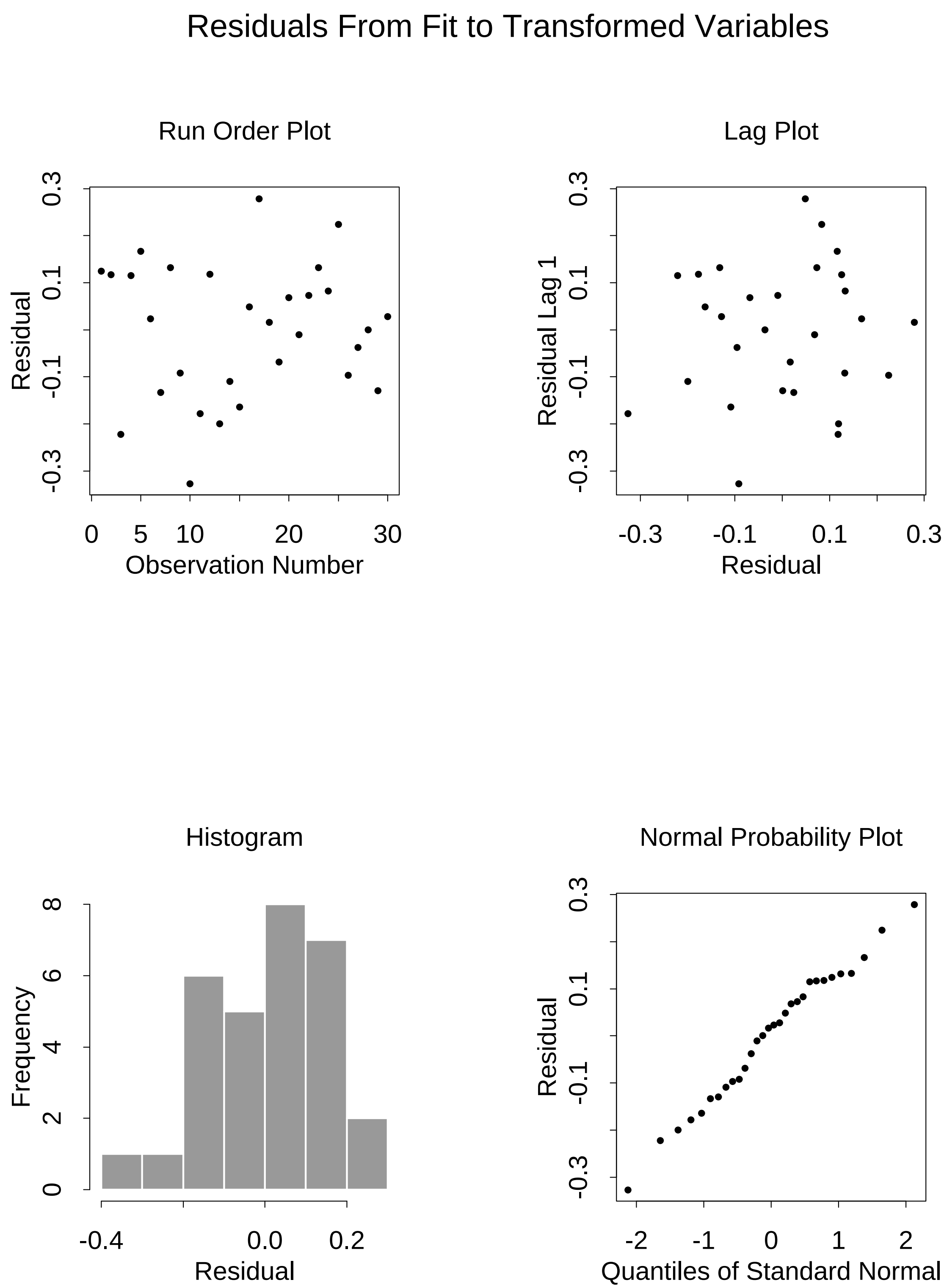



Transformed Response vs Predictor in Original Units
log(y)
7.0
6.5
6.0
5.5
5.0
20
40
60
80
100
x


Transformed Response vs Basic Tranformations
of the Predictor Variable
log(y)
5.0
6.0
7.0
20
40
60
80
100
x
log(y)
5.0
6.0
7.0
4
6
8
10
sqrt(x)
log(y)
5.0
6.0
7.0
2.5
3.5
4.5
log(x)
log(y)
5.0
6.0
7.0
0.02
0.06
0.10
1/x


Transformed Response vs Transformed Predictor
log(y)
7.0
6.5
6.0
5.5
5.0
2.5
3.0
3.5
4.0
4.5
log(x)


Residuals From Fit to Transformed Variables
Residuals
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
2.5
3.0
3.5
4.0
4.5
log(x)


Residuals From Fit to Transformed Variables
Run Order Plot
Residual
-0.3
-0.1
0.1
0.3
0
5
10
20
30
Observation Number
Lag Plot
Residual Lag 1
-0.3
-0.1
0.1
0.3
Residual
Histogram
Frequency
0
2
4
6
8
-0.4
0.0
0.2
Residual
Normal Probability Plot
Residual
-2
-1
0
1
2
Quantiles of Standard Normal

Weighted Least Squares

Weighted least squares estimates are found by minimizing:

$$Q = \sum_{i=1}^{n} w_i [y_i - f(x_{1i}, x_{2i}, \ldots, x_{ki}; \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)]^2$$

where

$$w_i \propto \frac{1}{\sigma_i^2}$$

These relative weights give optimal results, when known.

Weighted Least Squares

Unfortunately the true weights are rarely known, so they have to be estimated.

The obvious way to estimate the w_i, if there are replicates in the data, is

$$w_i = \frac{1}{\hat{\sigma}_i^2} = \frac{1}{s_i^2} = [\frac{\Sigma_{i=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n_i - 1}]^{-1}$$

for the i^{th} set of replicates in the data set.

However, this rarely works well because the weights are extremely variable when estimate this way.

Weighted Least Squares

A better strategy for estimating the weights is to find a function which relates s_i^2 to x_i.

If

$$s_i^2 \propto f(x_{1i}, x_{2i}, \ldots, x_{ki})$$

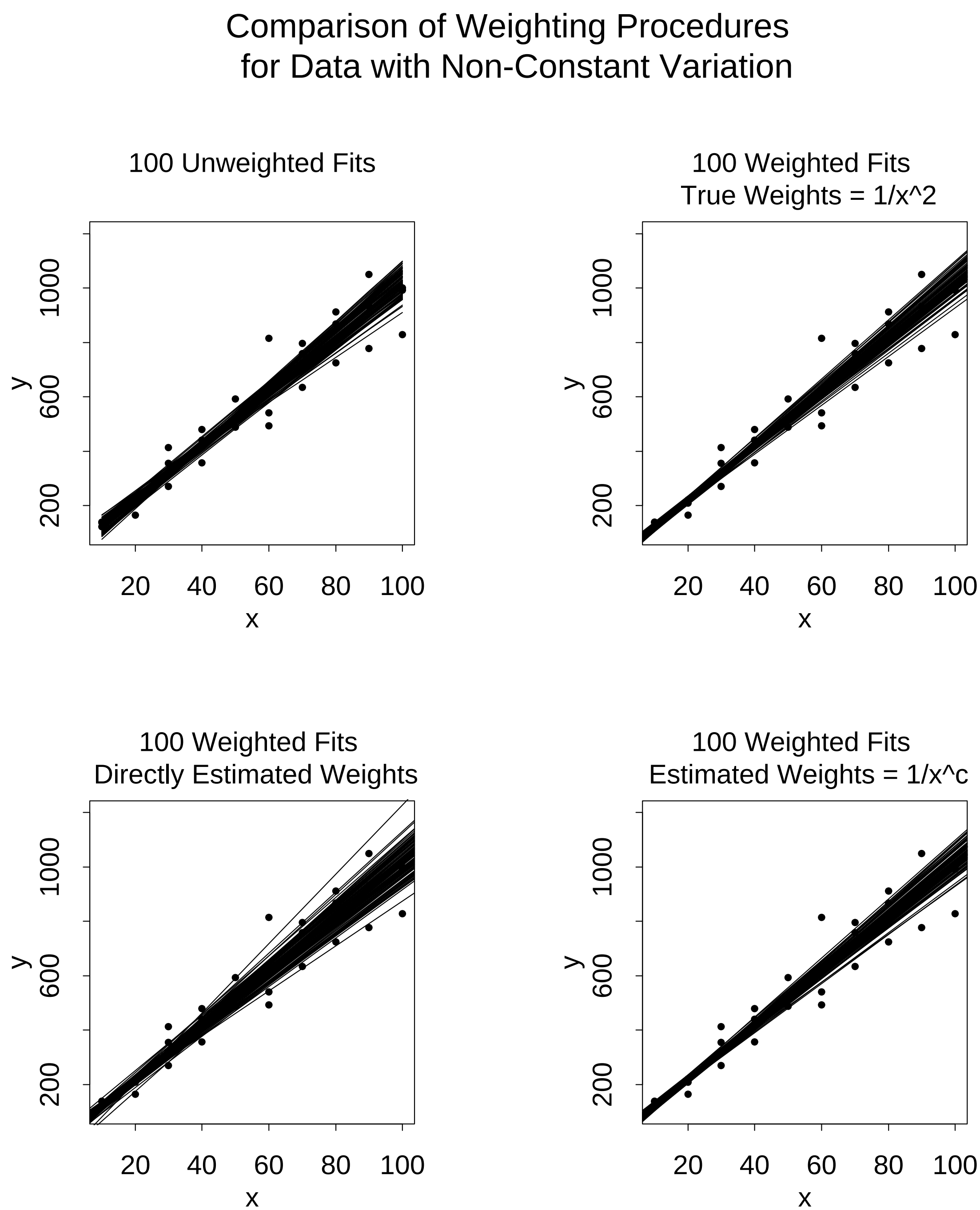
then use

$$w_i = \frac{1}{f(x_{1i}, x_{2i}, \ldots, x_{ki})}$$

as the weights.

One model that *often* works well for modeling the variances is

$$s_i^2 \propto x_i^c$$


Comparison of Weighting Procedures
for Data with Non-Constant Variation
100 Unweighted Fits
100 Weighted Fits
True Weights = 1/x^2
100 Weighted Fits
Directly Estimated Weights
100 Weighted Fits
Estimated Weights = 1/x^c
y
x
200
600
1000
20
40
60
80
100

Estimation of the Weights - Replicates Available

To estimate the weights using the power function shown above, fit the function

$$\log(s_i^2) = \beta_1 + \beta_2 \log(x_i)$$

to the variances from each set of replicates in the data.

The use $\hat{c} = \hat{\beta}_2$ (the slope of the fit) to estimate c, and

$$w_i = \frac{1}{x_i^{\hat{c}}}$$

to as the weights.

Check the residuals from the fit used to estimate c just to make sure everything looks reasonable. The fit does not have to meet the standards usually used, however.

Estimation of the Weights - No Replicates Available

If there are few or no replicates in the data, divide the data into several ranges in which the responses have fairly similar means.

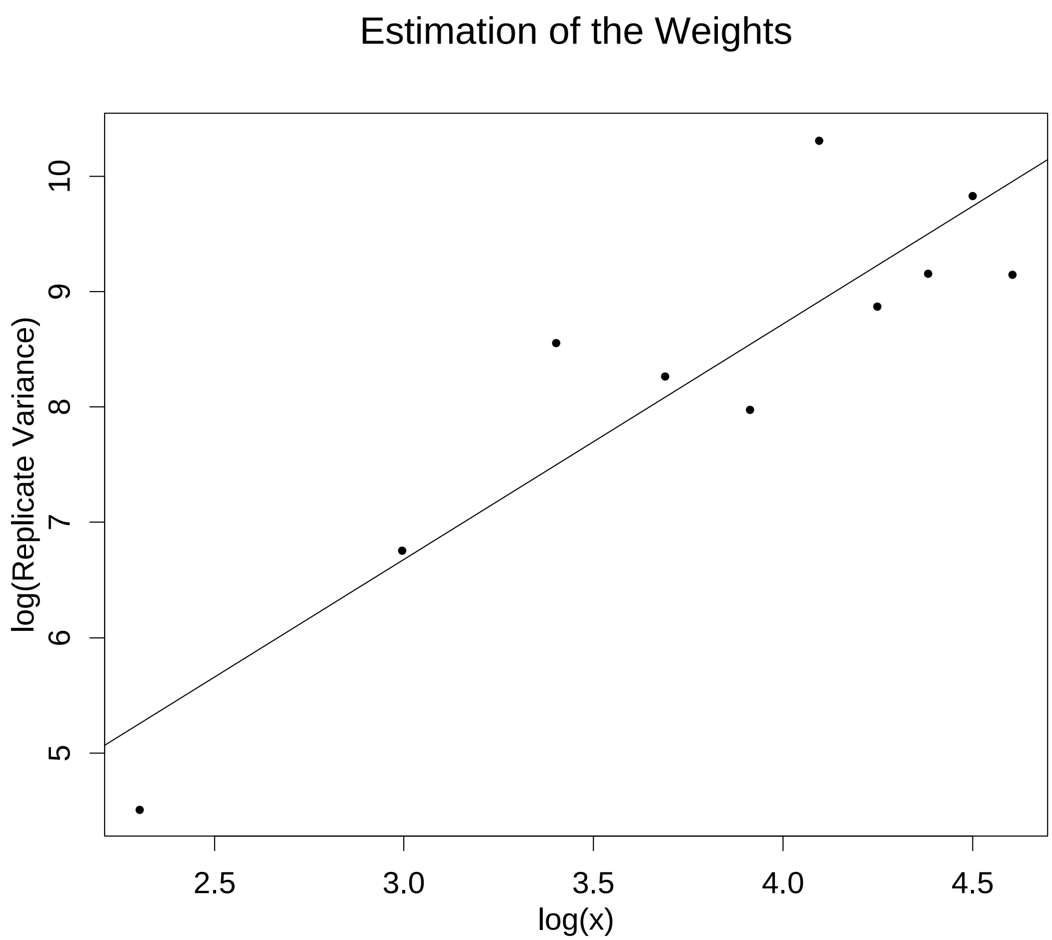
Treat each range as replicates and compute $\bar{x}_i$ and s_i^2 for each range.

Then fit

$$\log(s_i^2) = \beta_1 + \beta_2 \log(\bar{x}_i)$$

and define the weights by

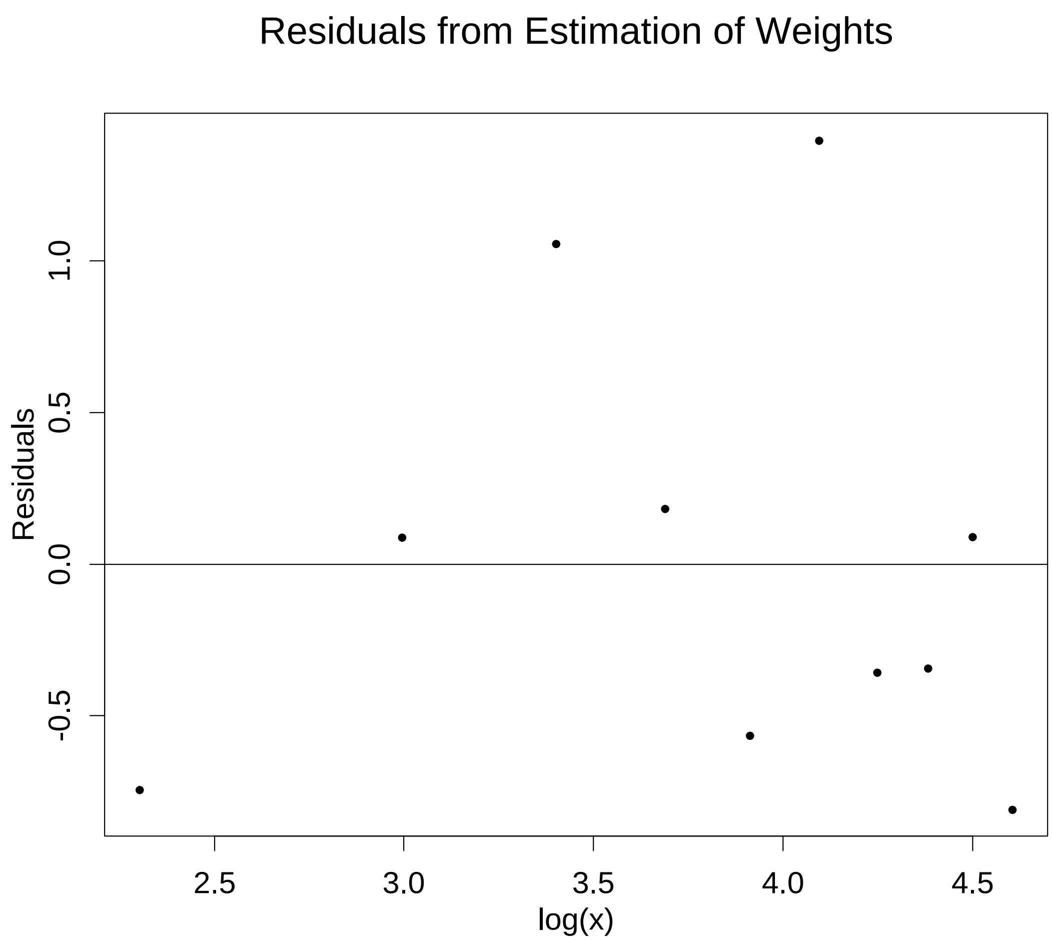
$$w_i = \frac{1}{x_i^{\hat{\beta}_2}}$$

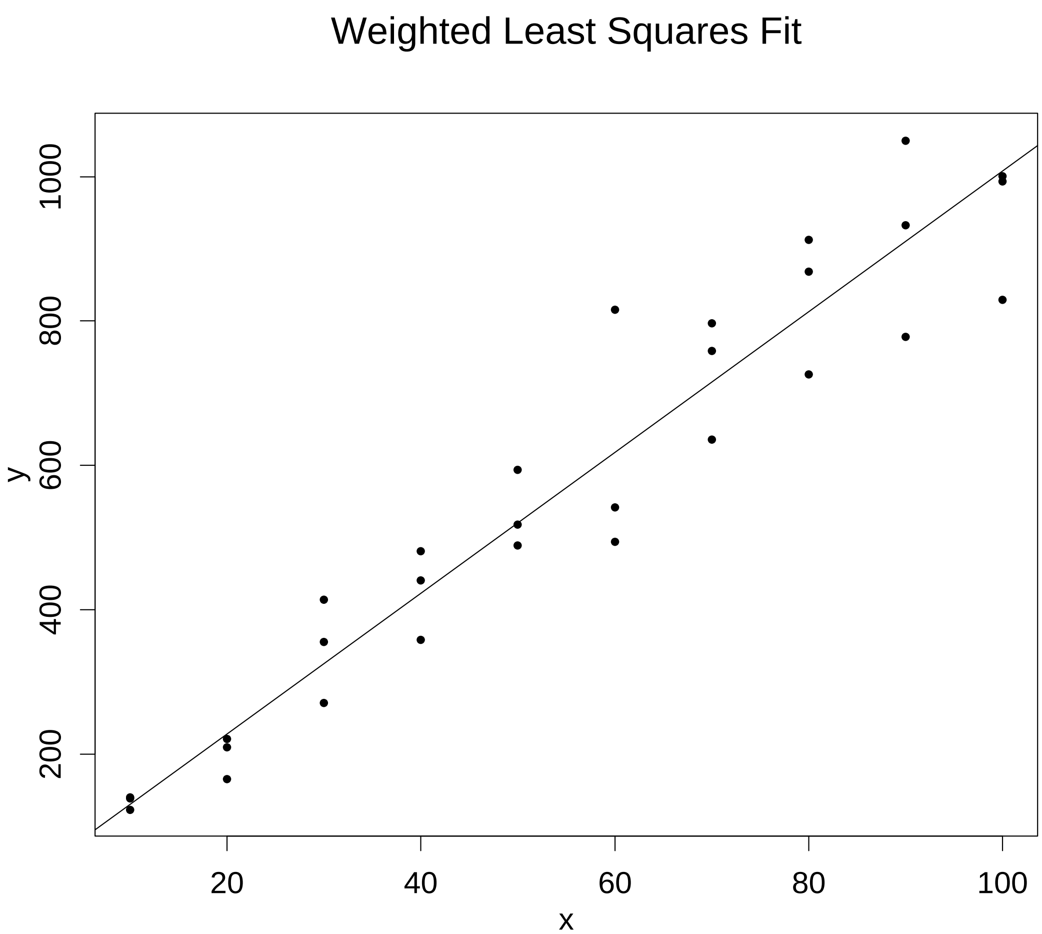


Estimation of the Weights
log(Replicate Variance)
5
6
7
8
9
10
2.5
3.0
3.5
4.0
4.5
log(x)

Output from Fit for Weight Estimation

```
N = 10

Residual Standard Error = 0.7821

Multiple R-Square = 0.8046

F-statistic = 32.9428 on 1 and 8 df, p-value = 4e-04

             coef std.err  t.stat p.value
Intercept  0.5554  1.3785  0.4029  0.6976
        X  2.0414  0.3557  5.7396  0.0004
```


Residuals from Estimation of Weights
Residuals
1.0
0.5
0.0
-0.5
2.5
3.0
3.5
4.0
4.5
log(x)


Weighted Least Squares Fit
y
1000
800
600
400
200
x
20
40
60
80
100

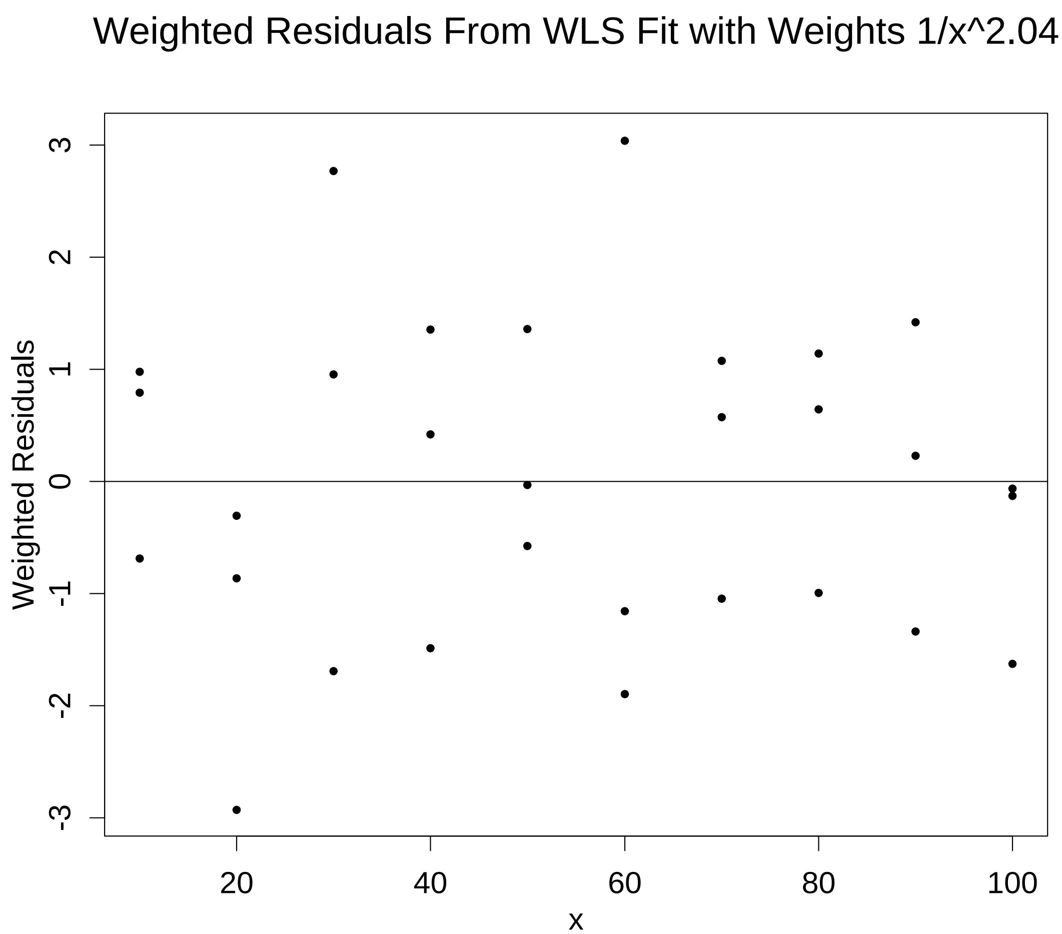
Weighted Residuals

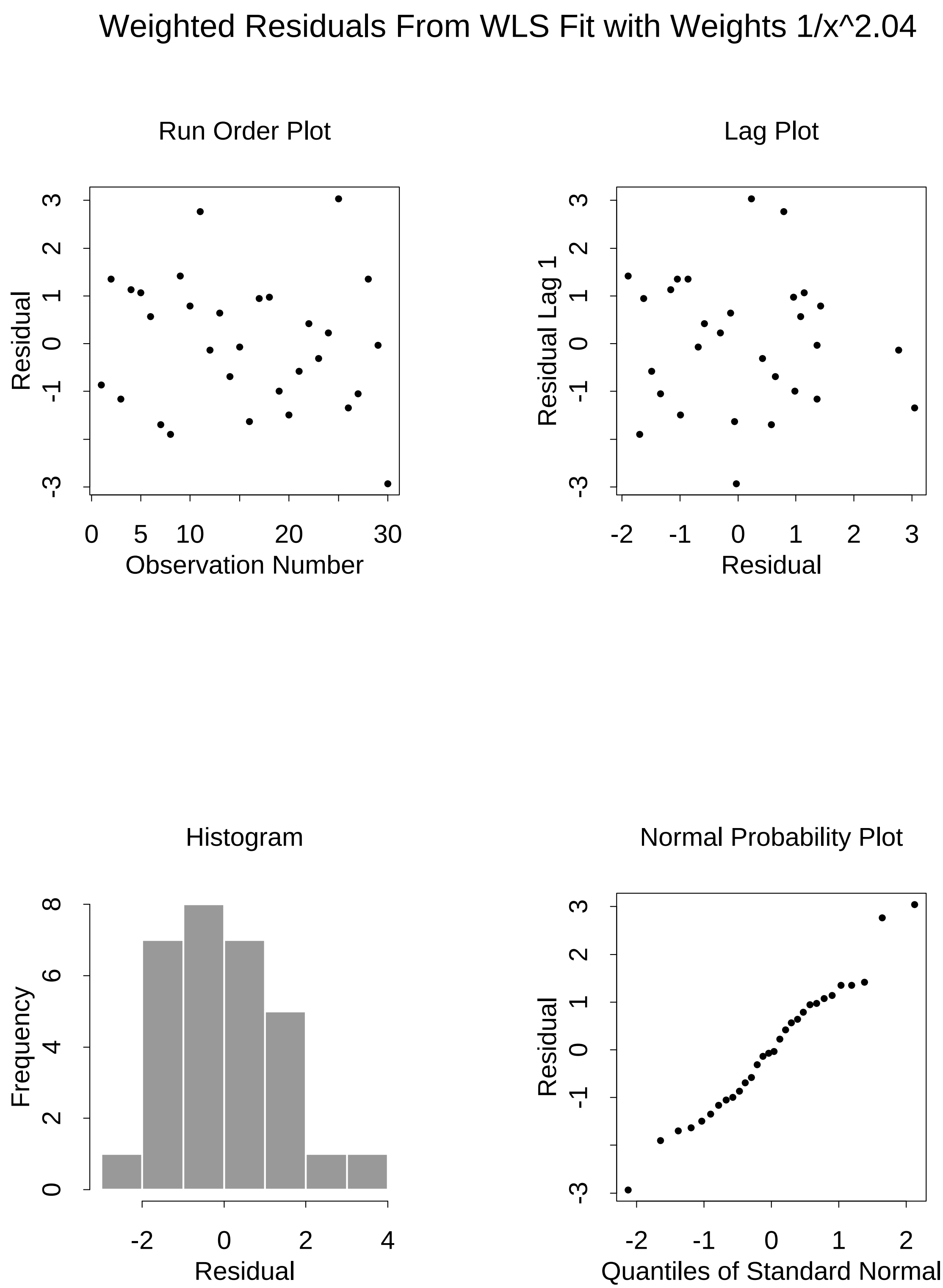
One complication with weighted analyses is the fact that the distribution of the residuals can vary substantially with the different values of the predictor variables.

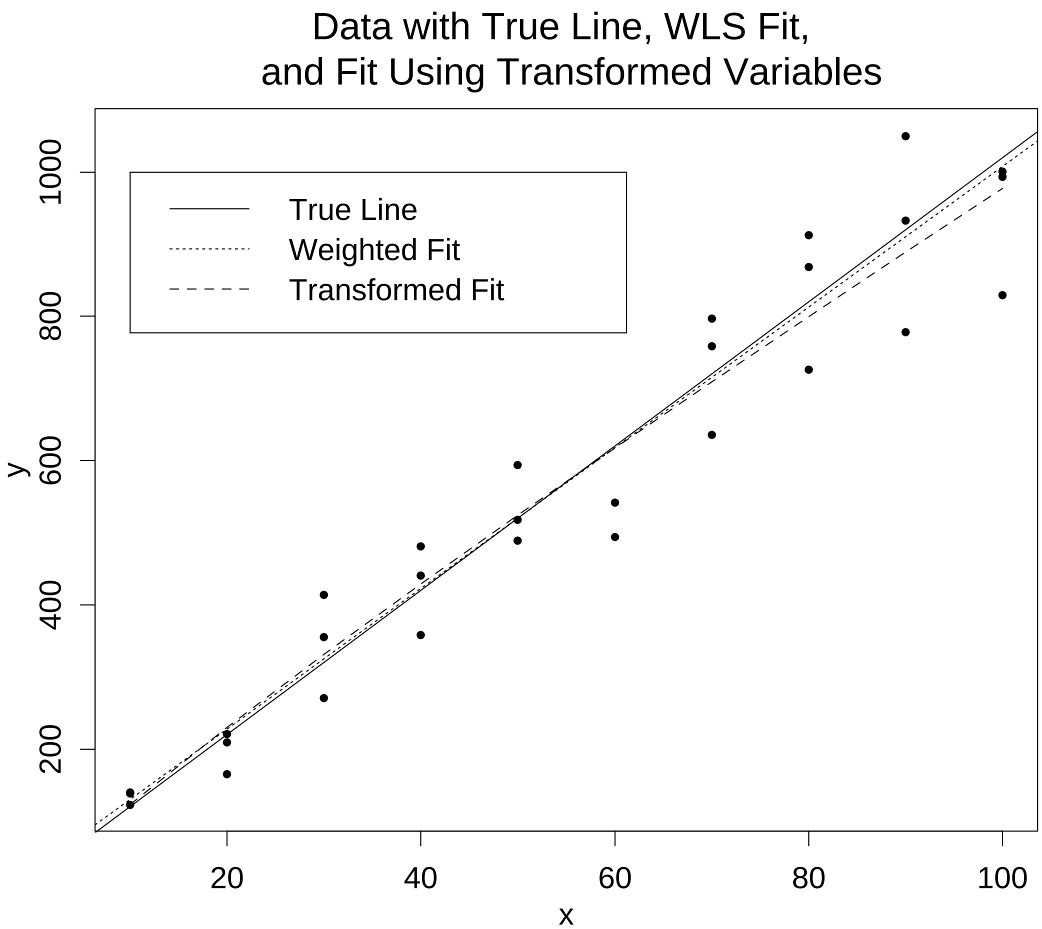
This necessitates the use of weighted residuals when plotting residuals.

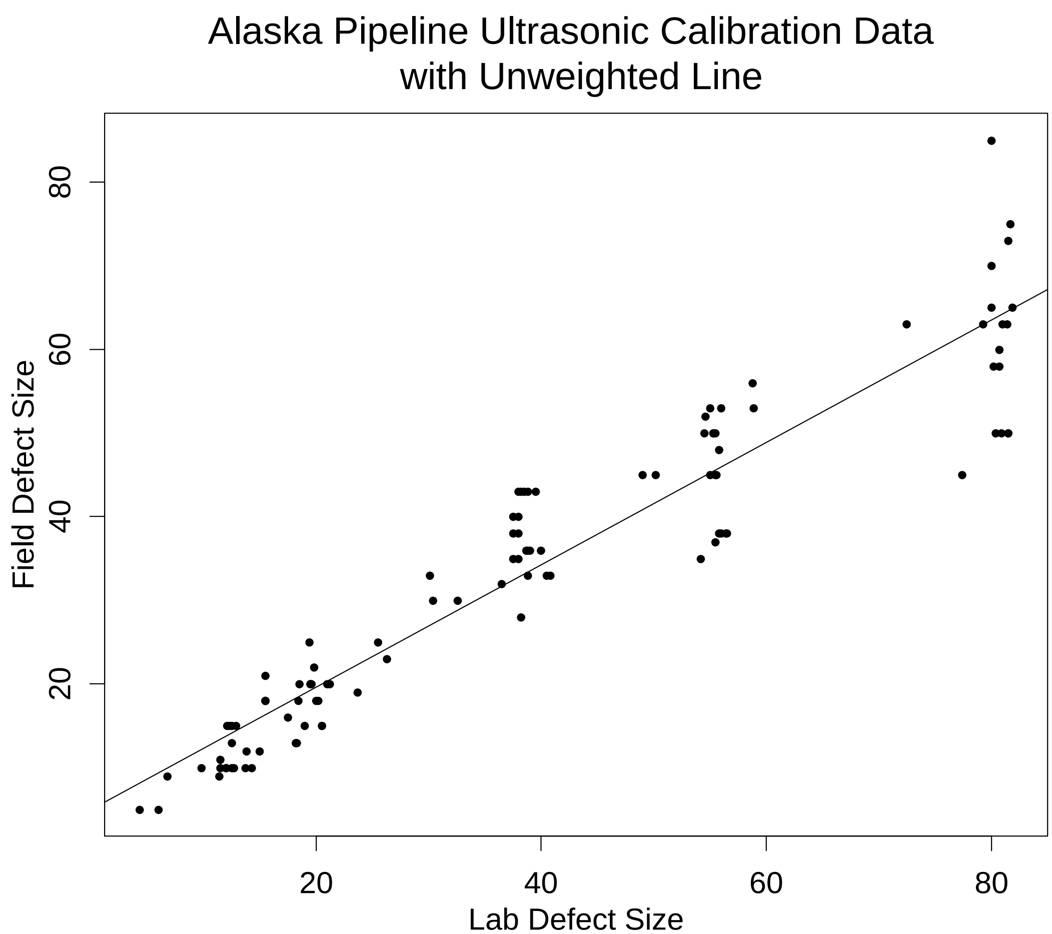
The weighted residuals are given by

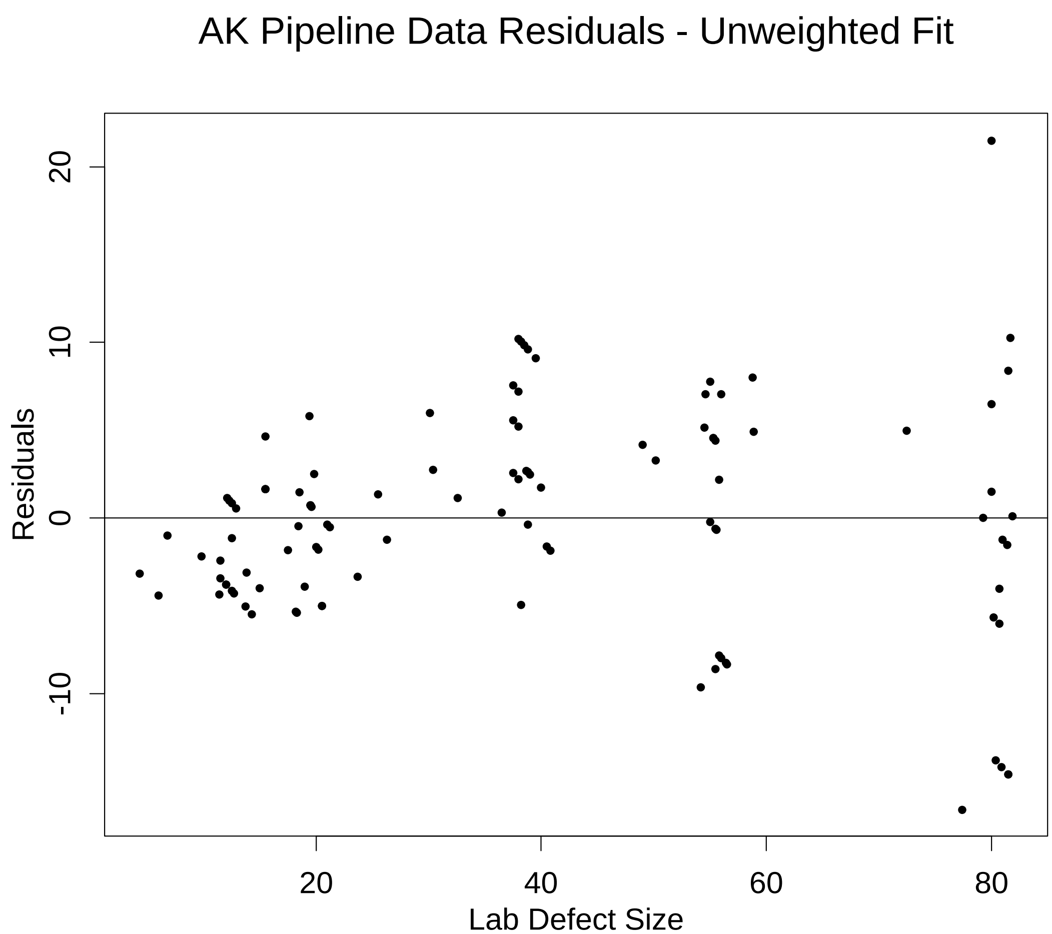
$$e_i = \sqrt{w_i}(y_i - f(x_{1i}, \ldots, x_{ki}; \hat{\beta}_1, \ldots, \hat{\beta}_p))$$


Weighted Residuals From WLS Fit with Weights 1/x^2.04
Weighted Residuals
3
2
1
0
-1
-2
-3
20
40
60
80
100
x


Weighted Residuals From WLS Fit with Weights 1/x^2.04
Run Order Plot
Residual
Observation Number
Lag Plot
Residual Lag 1
Residual
Histogram
Frequency
Residual
Normal Probability Plot
Residual
Quantiles of Standard Normal

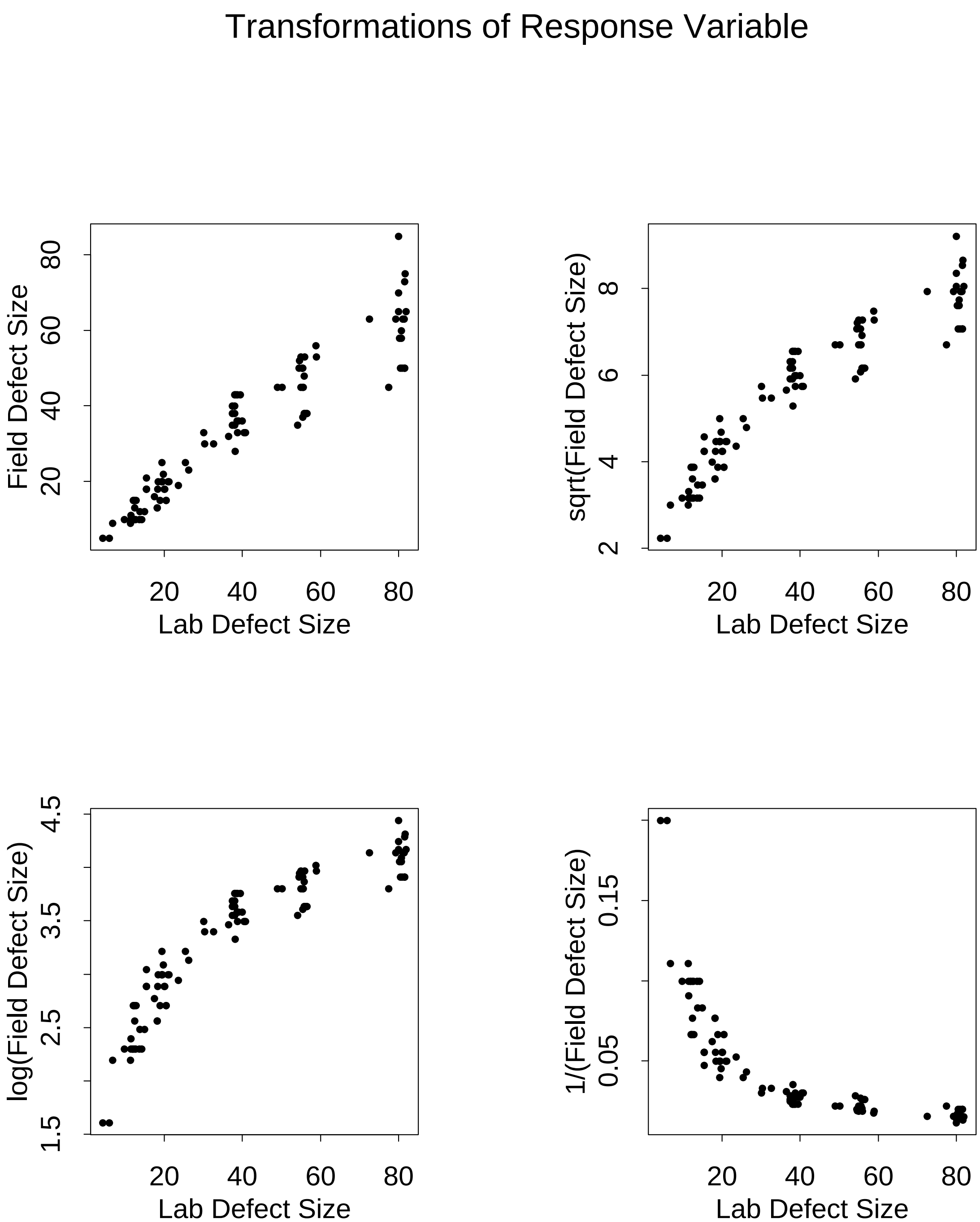

Data with True Line, WLS Fit,
and Fit Using Transformed Variables
True Line
Weighted Fit
Transformed Fit
y
x
200
400
600
800
1000
20
40
60
80
100

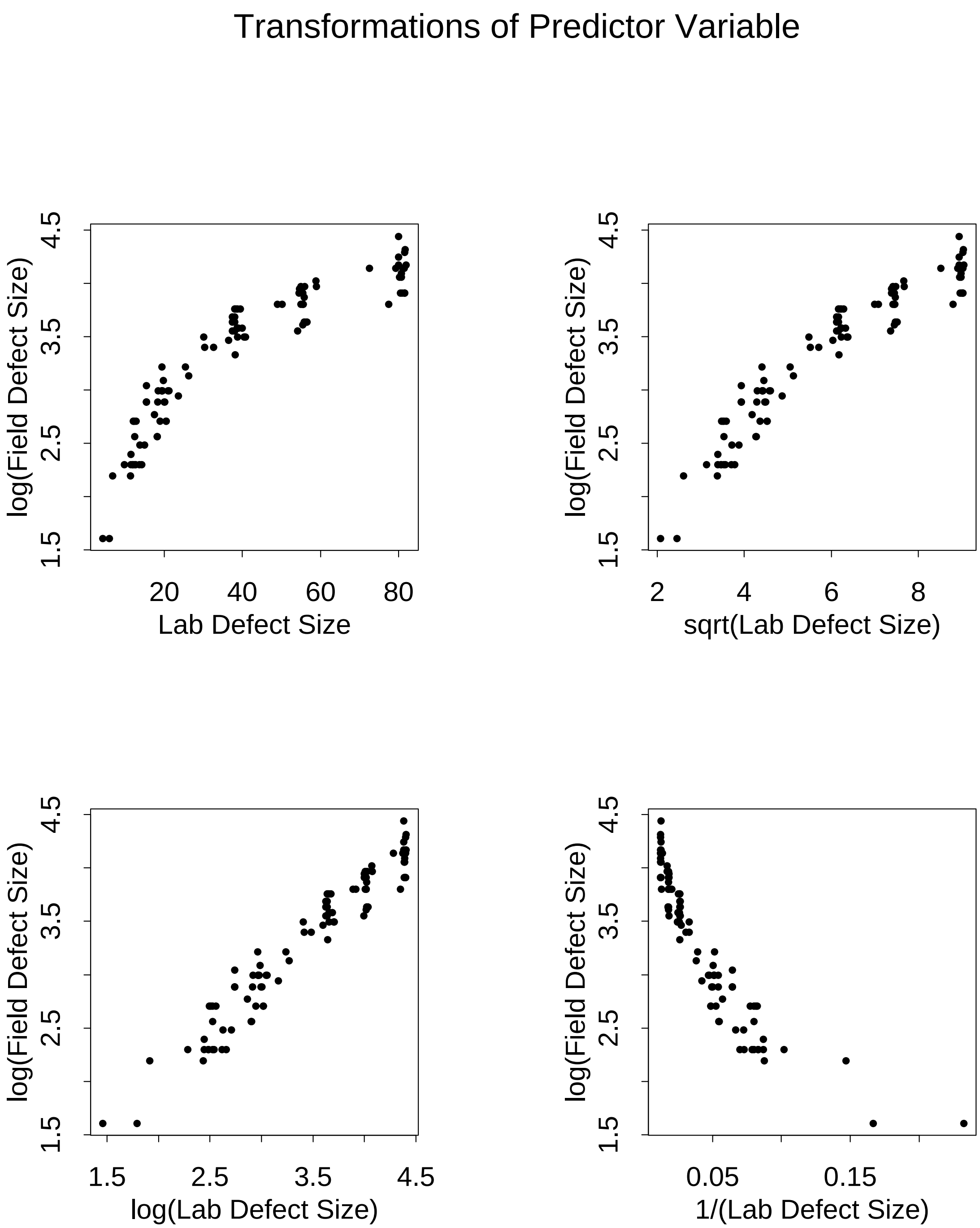

Alaska Pipeline Ultrasonic Calibration Data
with Unweighted Line
Field Defect Size
Lab Defect Size
20
40
60
80
20
40
60
80

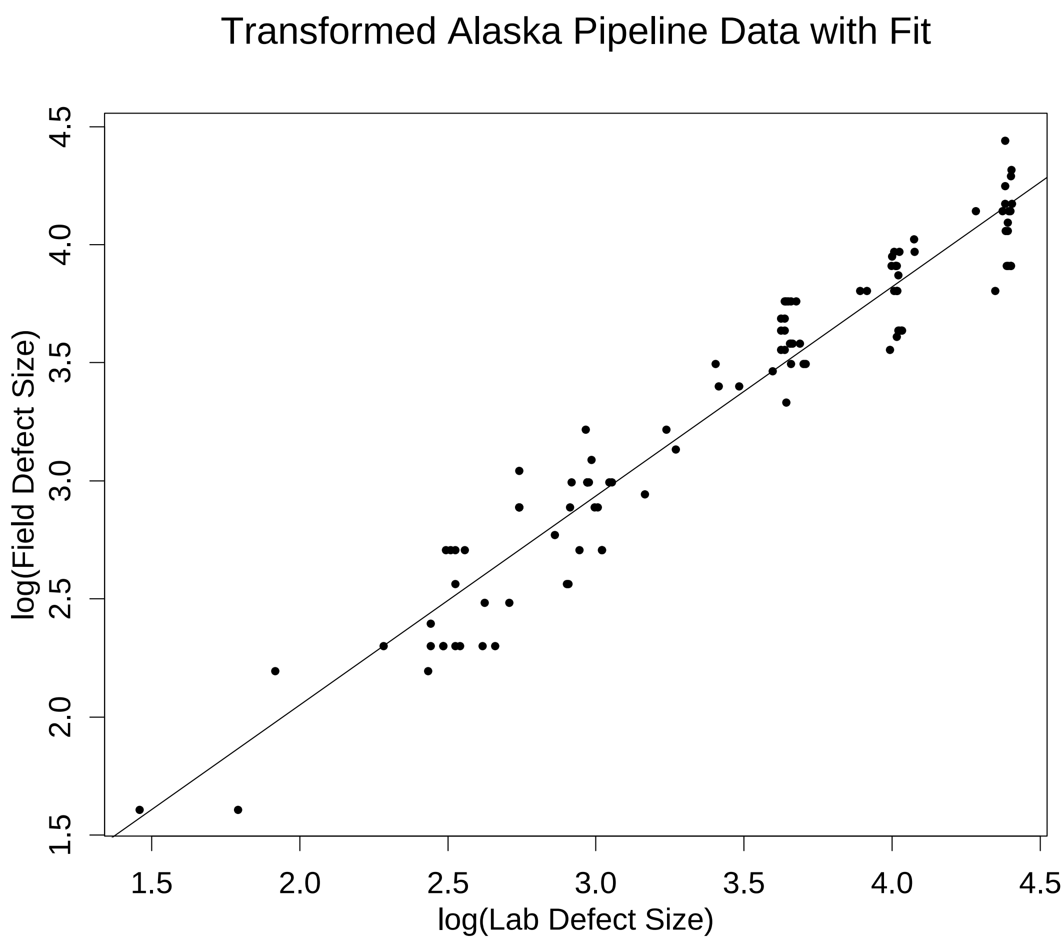

AK Pipeline Data Residuals - Unweighted Fit
Residuals
Lab Defect Size
20
10
0
-10
20
40
60
80

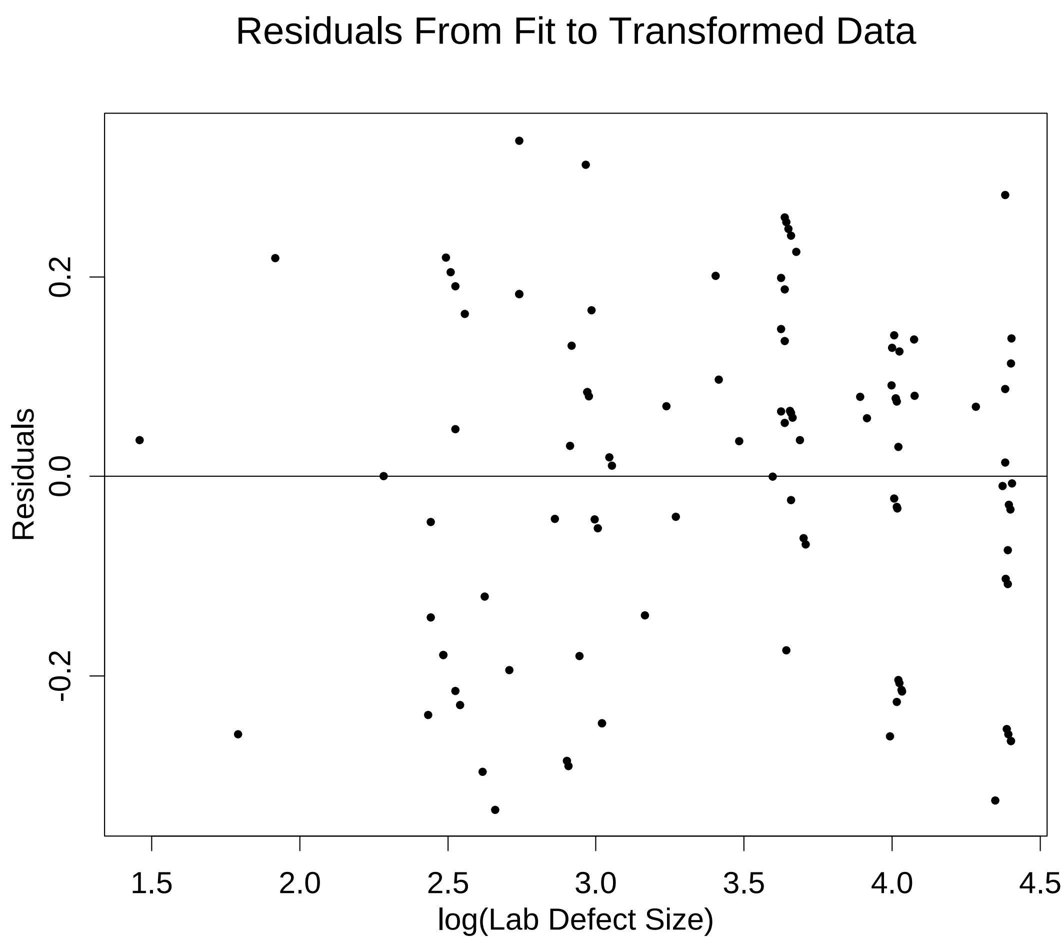
AK Pipeline Data
Output from Unweighted Fit

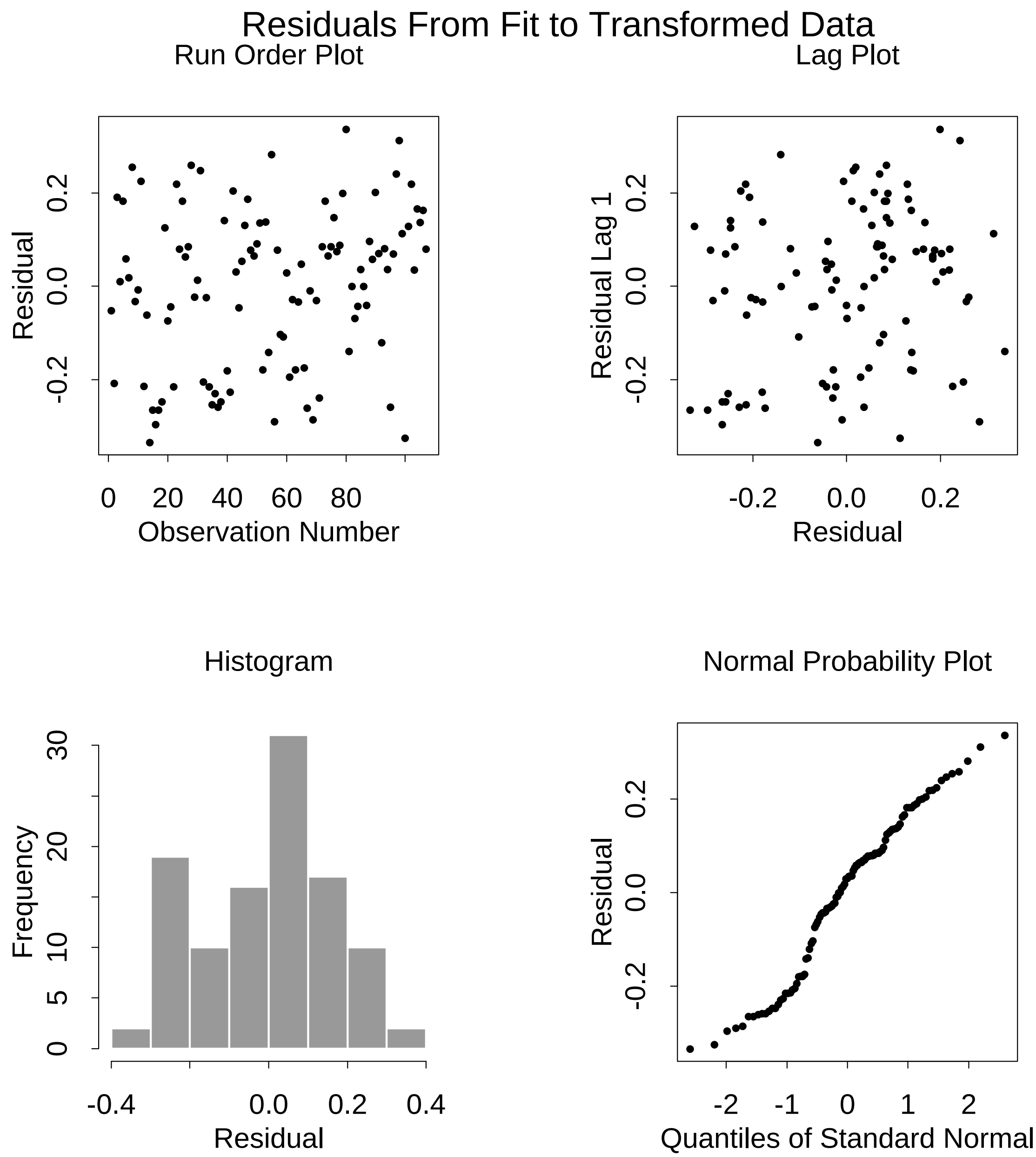
```
N = 107

Residual Standard Error = 6.080924

Multiple R-Square = 0.8941251

F-statistic = 886.7366 on 1 and 105 df, p-value = 0

               coef    std.err    t.stat      p.value
Intercept 4.9936799 1.12565780  4.436233 2.263517e-05
        X 0.7311111 0.02455195 29.778123 0.000000e+00
```


Transformations of Response Variable
Field Defect Size
20
40
60
80
Lab Defect Size
sqrt(Field Defect Size)
2
4
6
8
log(Field Defect Size)
1.5
2.5
3.5
4.5
1/(Field Defect Size)
0.05
0.15


Transformations of Predictor Variable
log(Field Defect Size)
1.5
2.5
3.5
4.5
20
40
60
80
Lab Defect Size
2
4
6
8
sqrt(Lab Defect Size)
1.5
2.5
3.5
4.5
log(Lab Defect Size)
0.05
0.15
1/(Lab Defect Size)

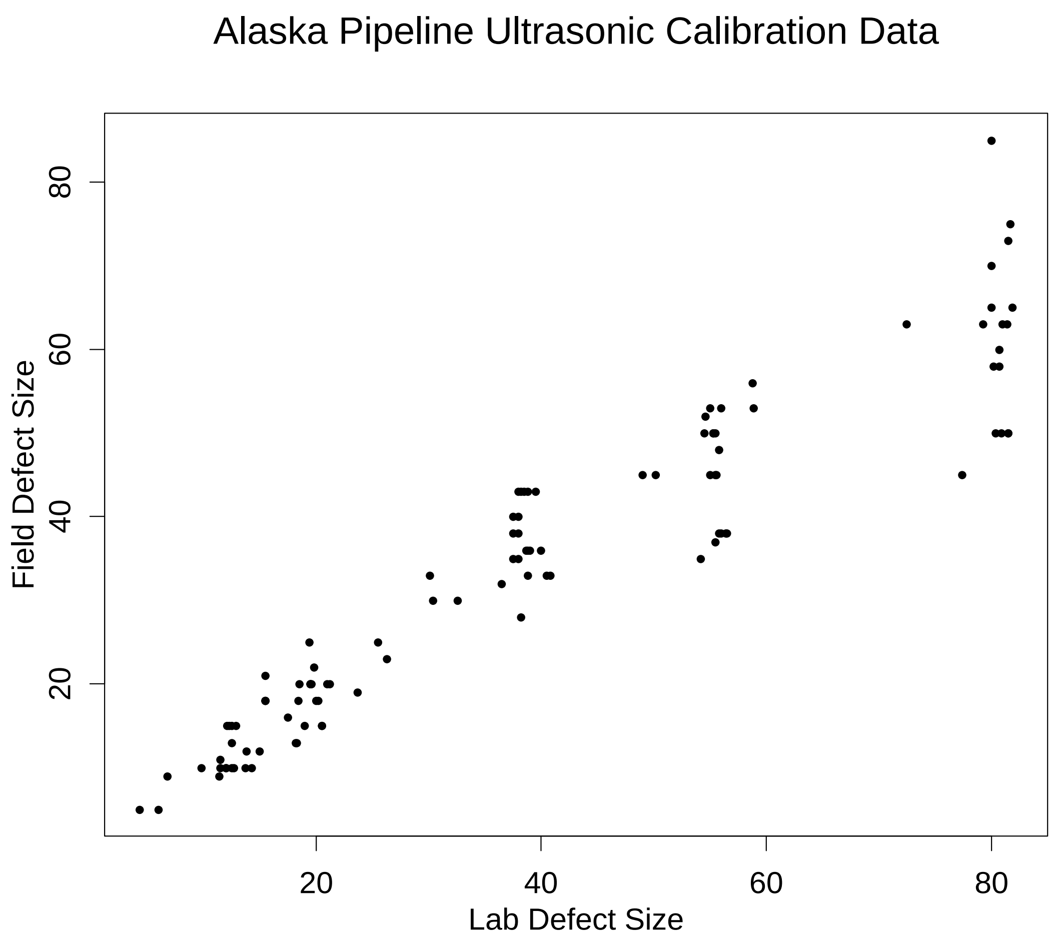

Transformed Alaska Pipeline Data with Fit
log(Field Defect Size)
log(Lab Defect Size)
1.5
2.0
2.5
3.0
3.5
4.0
4.5
1.5
2.0
2.5
3.0
3.5
4.0
4.5

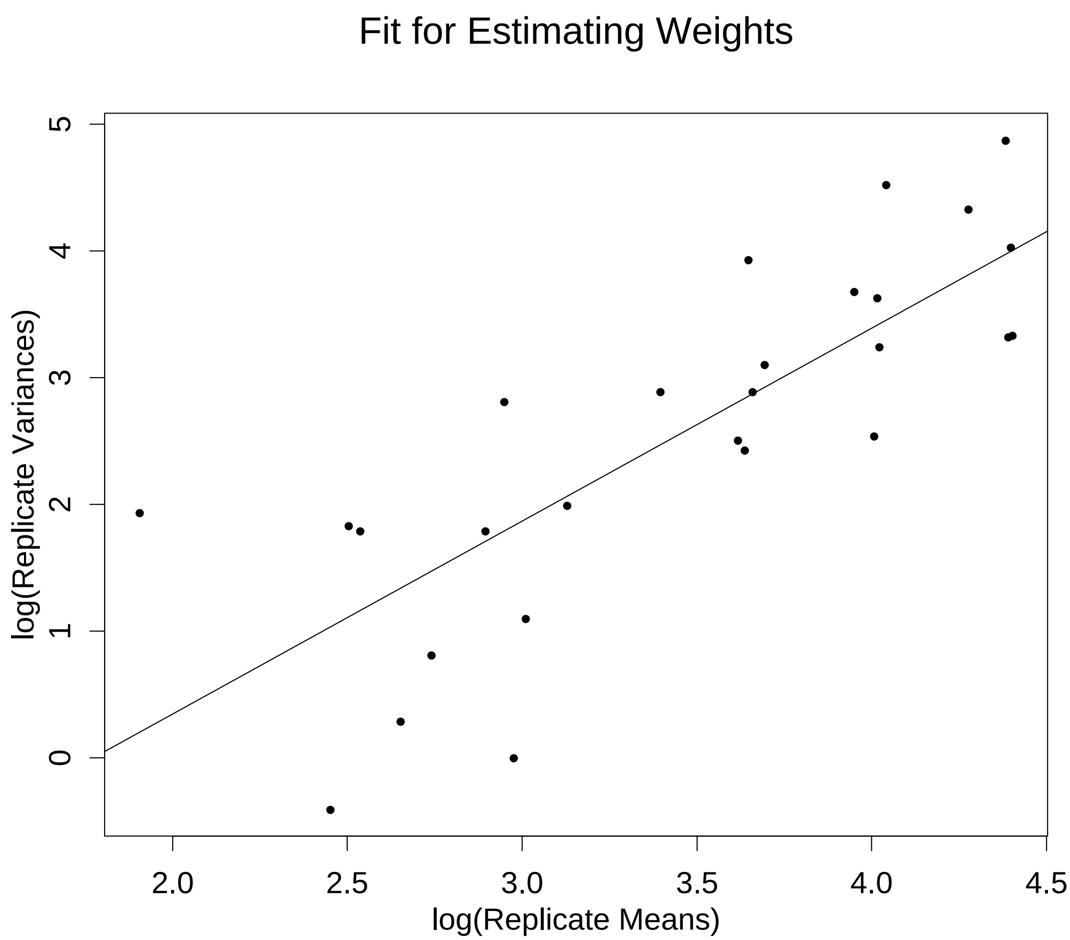

Residuals From Fit to Transformed Data
Residuals
0.2
0.0
-0.2
1.5
2.0
2.5
3.0
3.5
4.0
4.5
log(Lab Defect Size)

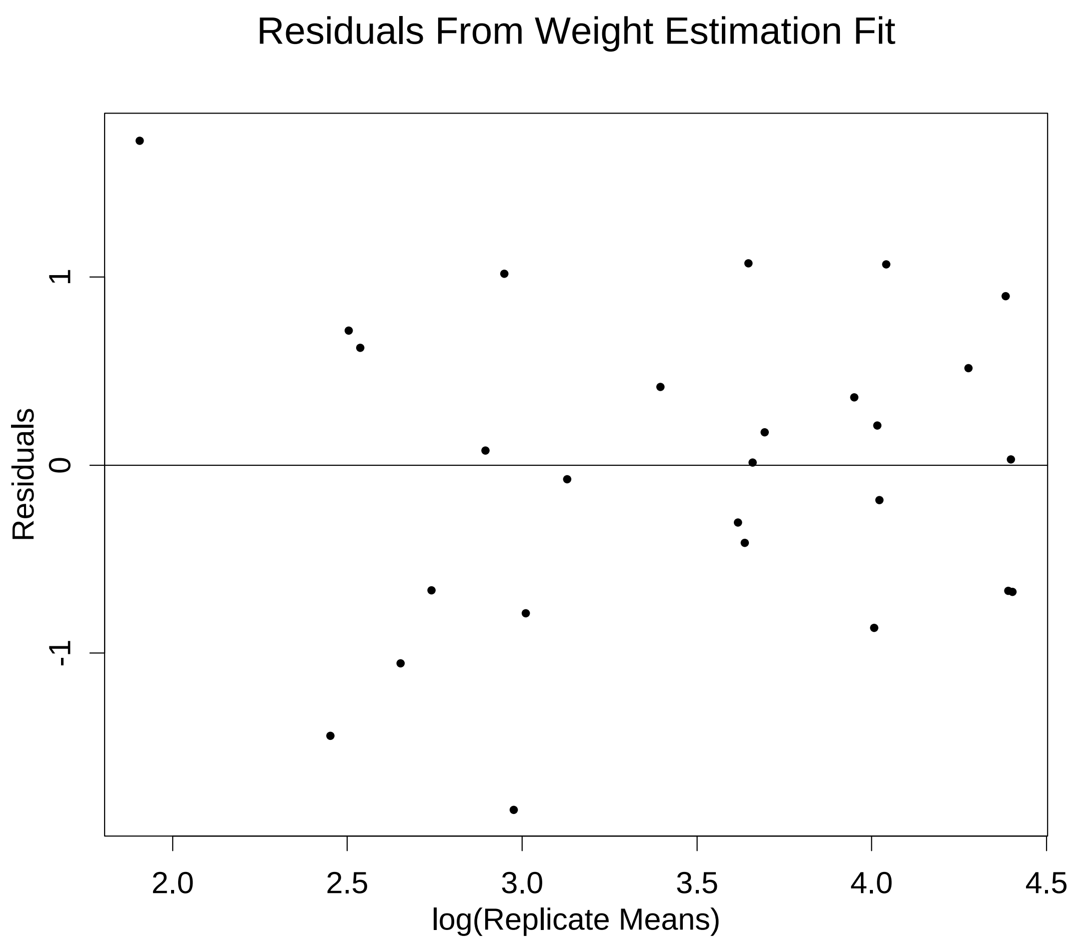

Residuals From Fit to Transformed Data
Run Order Plot
Residual
-0.2
0.0
0.2
0
20
40
60
80
Observation Number
Lag Plot
Residual Lag 1
-0.2
0.0
0.2
Residual
Histogram
Frequency
0
5
10
20
30
-0.4
0.0
0.2
0.4
Residual
Normal Probability Plot
Residual
-2
-1
0
1
2
Quantiles of Standard Normal

AK Pipeline Data
Output from Fit on Transformed Variables

```
N = 107

Residual Standard Error = 0.1682604

Multiple R-Square = 0.9337104

F-statistic = 1478.958 on 1 and 105 df, p-value = 0

               coef    std.err    t.stat      p.value
Intercept 0.2813838 0.08092894  3.476924 0.0007390395
        X 0.8851754 0.02301714 38.457221 0.0000000000
```

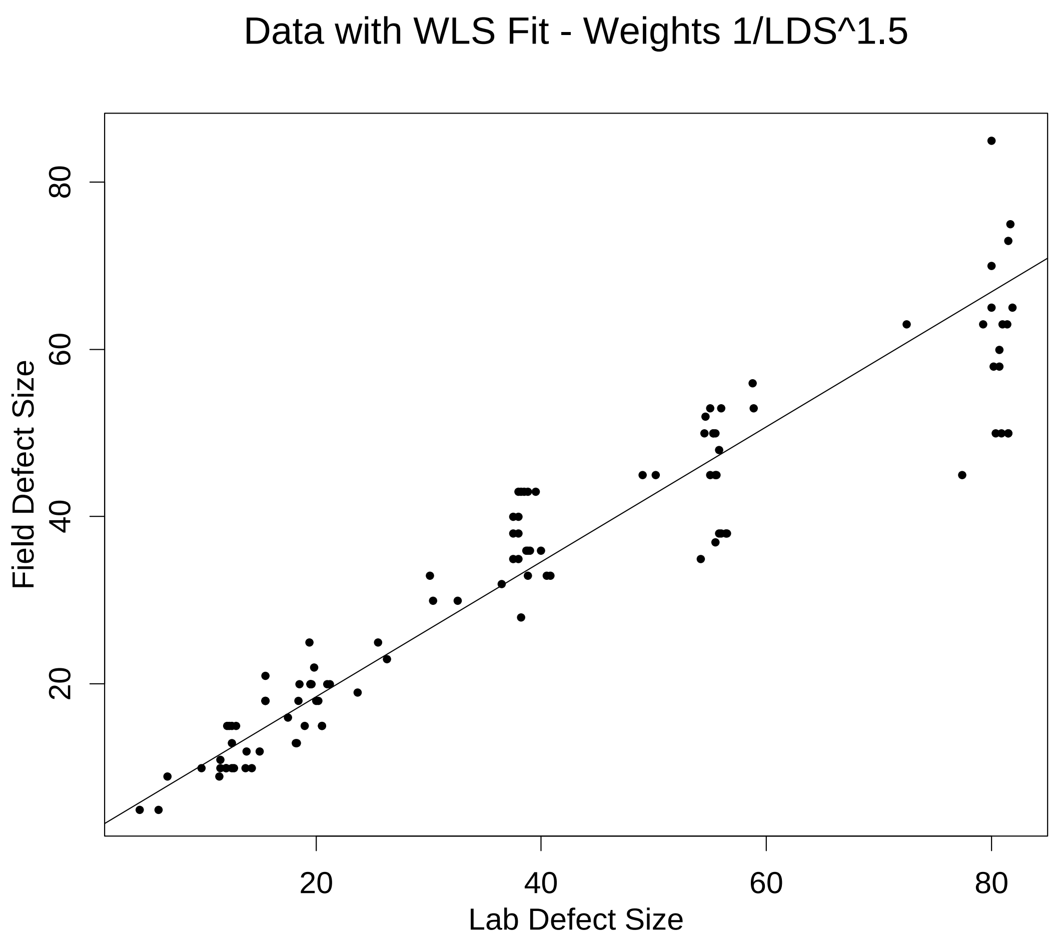

Alaska Pipeline Ultrasonic Calibration Data
Field Defect Size
Lab Defect Size
20
40
60
80
20
40
60
80

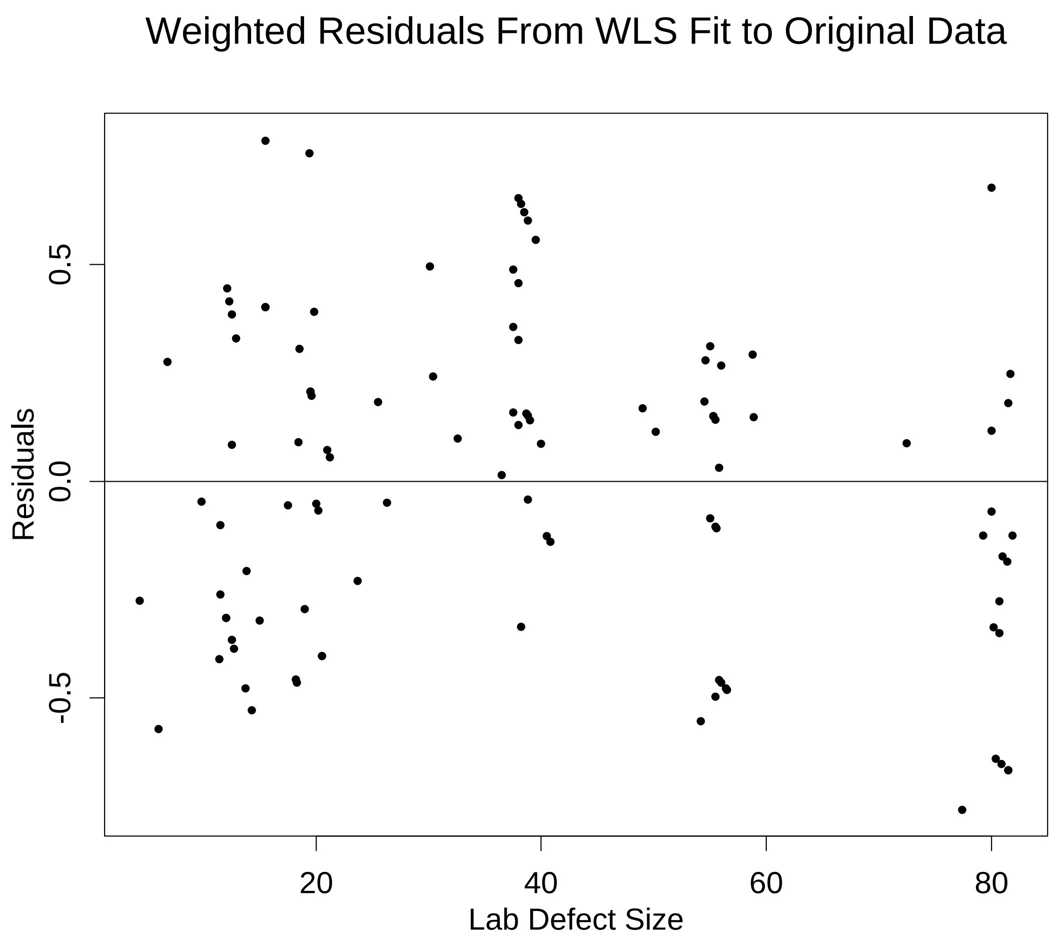

Fit for Estimating Weights
log(Replicate Variances)
log(Replicate Means)
0
1
2
3
4
5
2.0
2.5
3.0
3.5
4.0
4.5

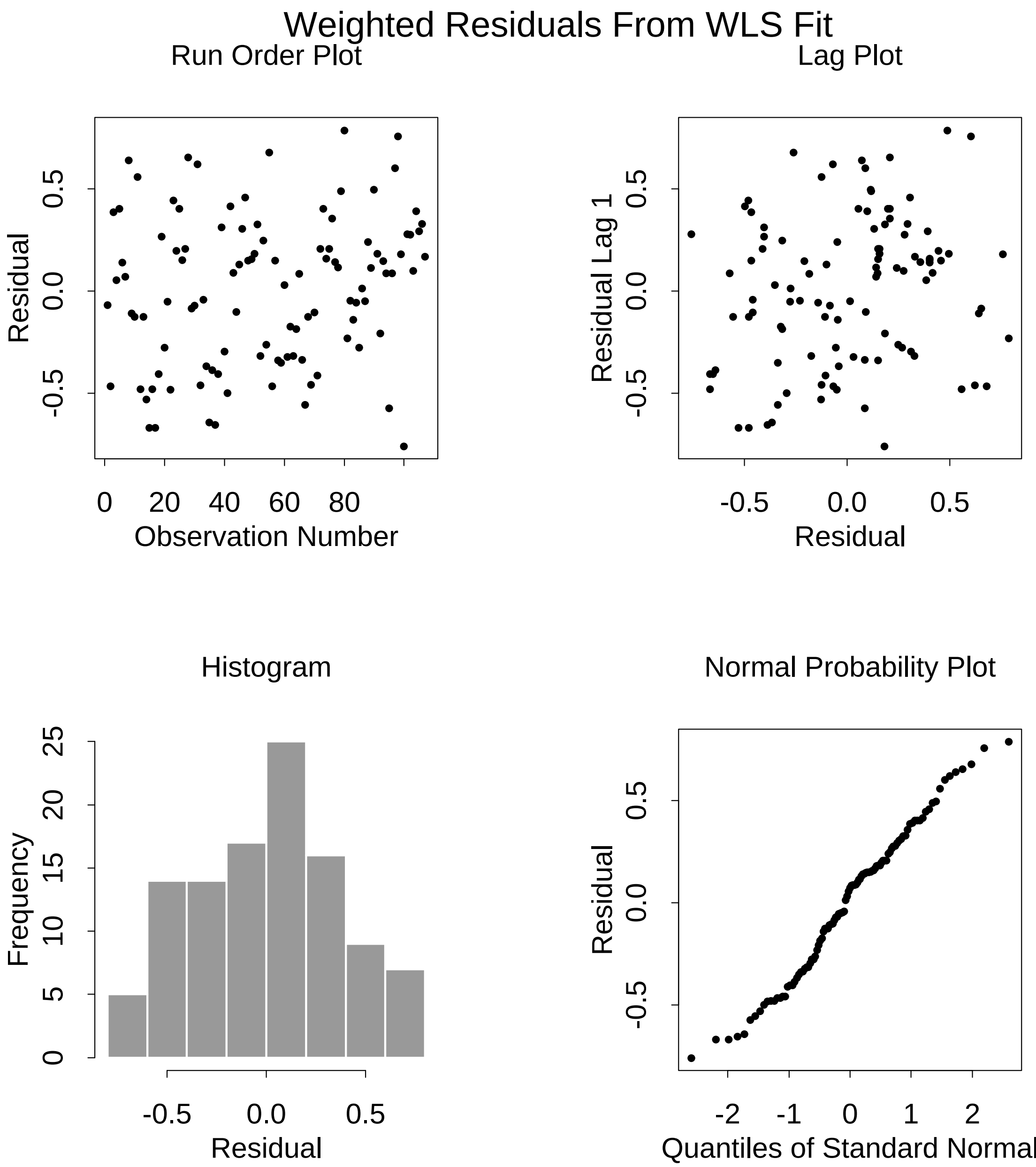

Residuals From Weight Estimation Fit
Residuals
1
0
-1
2.0
2.5
3.0
3.5
4.0
4.5
log(Replicate Means)

AK Pipeline Data
Output from Weight Estimation Fit

```
N = 27

Residual Standard Error = 0.8545392

Multiple R-Square = 0.6286482

F-statistic = 42.32161 on 1 and 25 df, p-value = 8.178602e-07

               coef   std.err    t.stat      p.value
Intercept -2.696496 0.8250090 -3.268444 3.140569e-03
        X  1.522101 0.2339712  6.505506 8.178602e-07
```

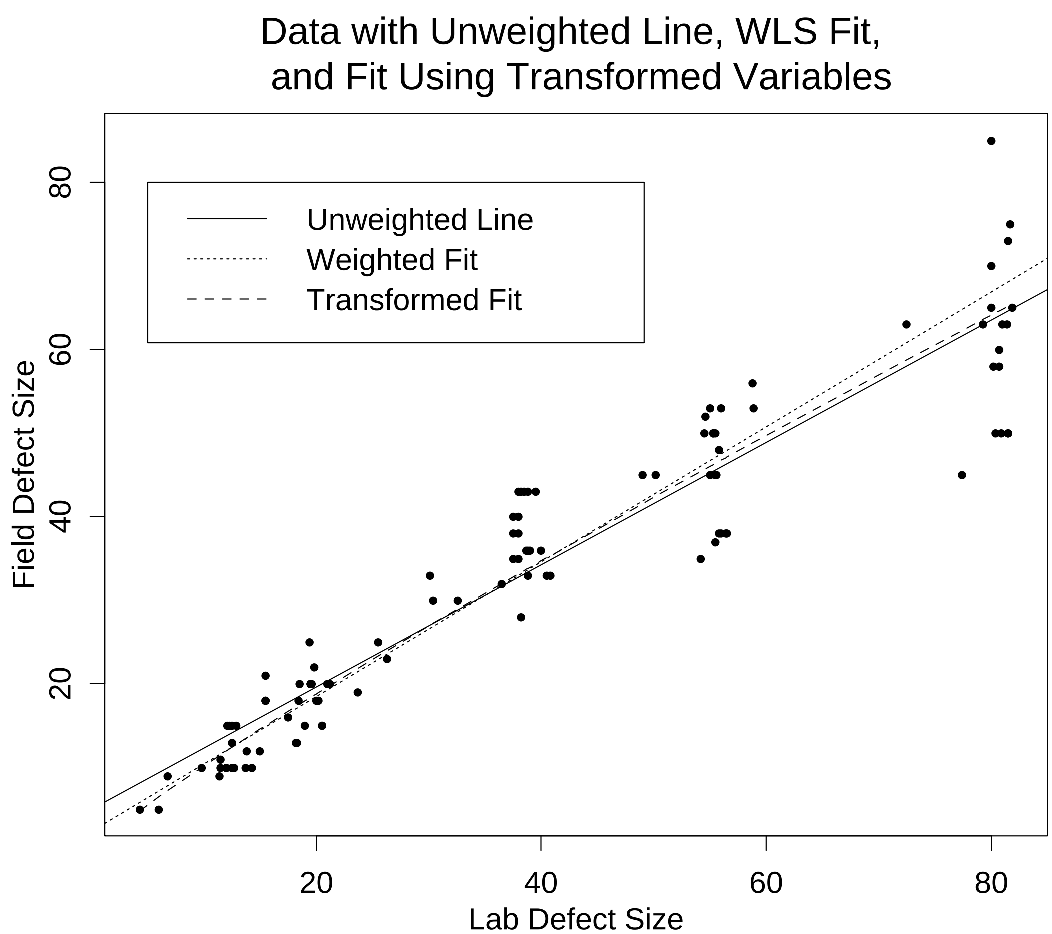

Data with WLS Fit - Weights 1/LDS^1.5
Field Defect Size
20
40
60
80
Lab Defect Size
20
40
60
80

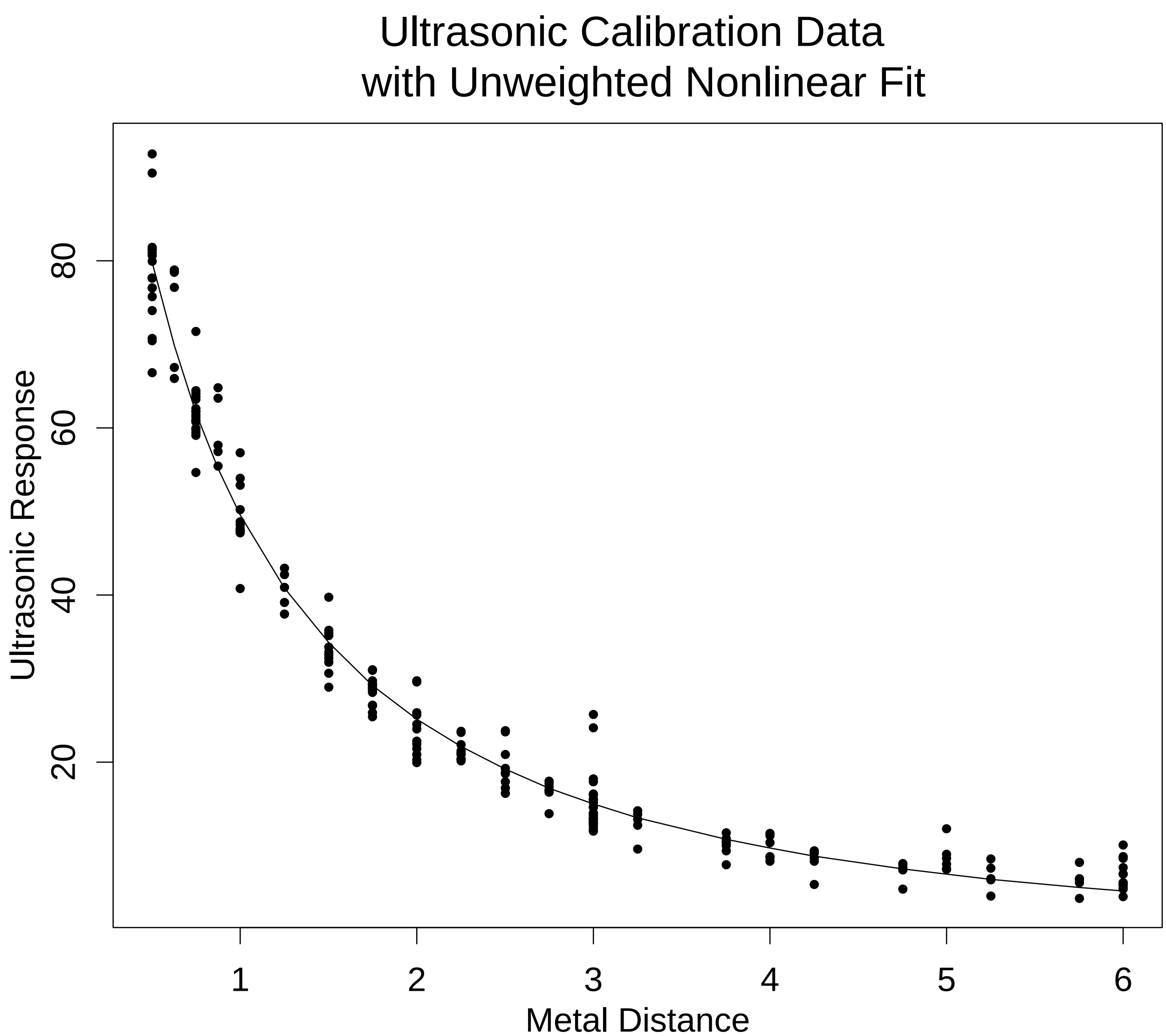

Weighted Residuals From WLS Fit to Original Data
Residuals
0.5
0.0
-0.5
20
40
60
80
Lab Defect Size

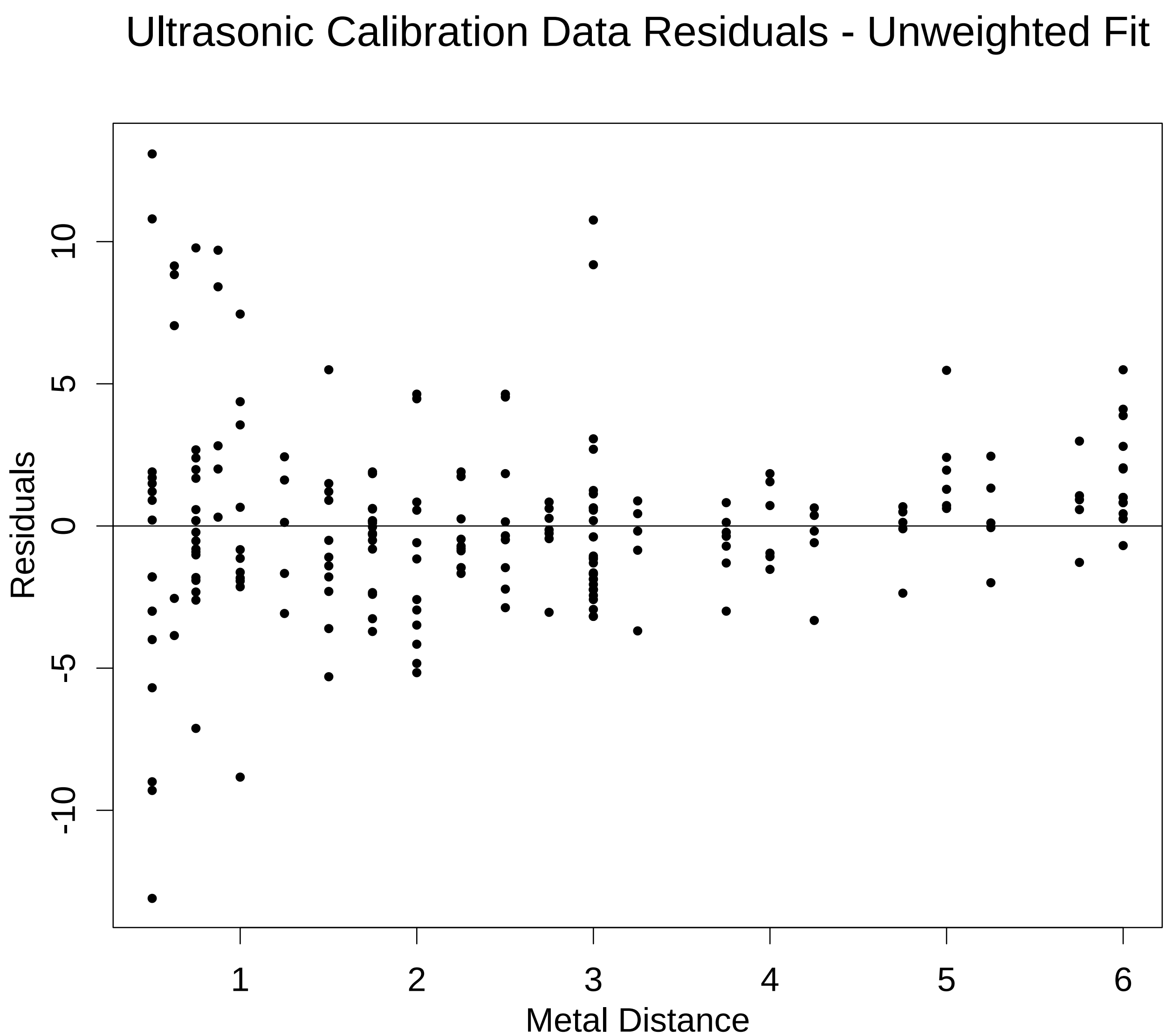

Weighted Residuals From WLS Fit
Run Order Plot
Residual
-0.5
0.0
0.5
0
20
40
60
80
Observation Number
Lag Plot
Residual Lag 1
-0.5
0.0
0.5
Residual
Histogram
Frequency
0
5
10
15
20
25
-0.5
0.0
0.5
Residual
Normal Probability Plot
Residual
-0.5
0.0
0.5
-2
-1
0
1
2
Quantiles of Standard Normal

AK Pipeline Data
Output from Weighted Fit

```
N = 107

Residual Standard Error = 0.3646

Multiple R-Square = 0.9235

F-statistic = 1267.024 on 1 and 105 df, p-value = 0

            coef std.err  t.stat p.value
Intercept 2.3523  0.5431  4.3312       0
        X 0.8064  0.0227 35.5953       0
```

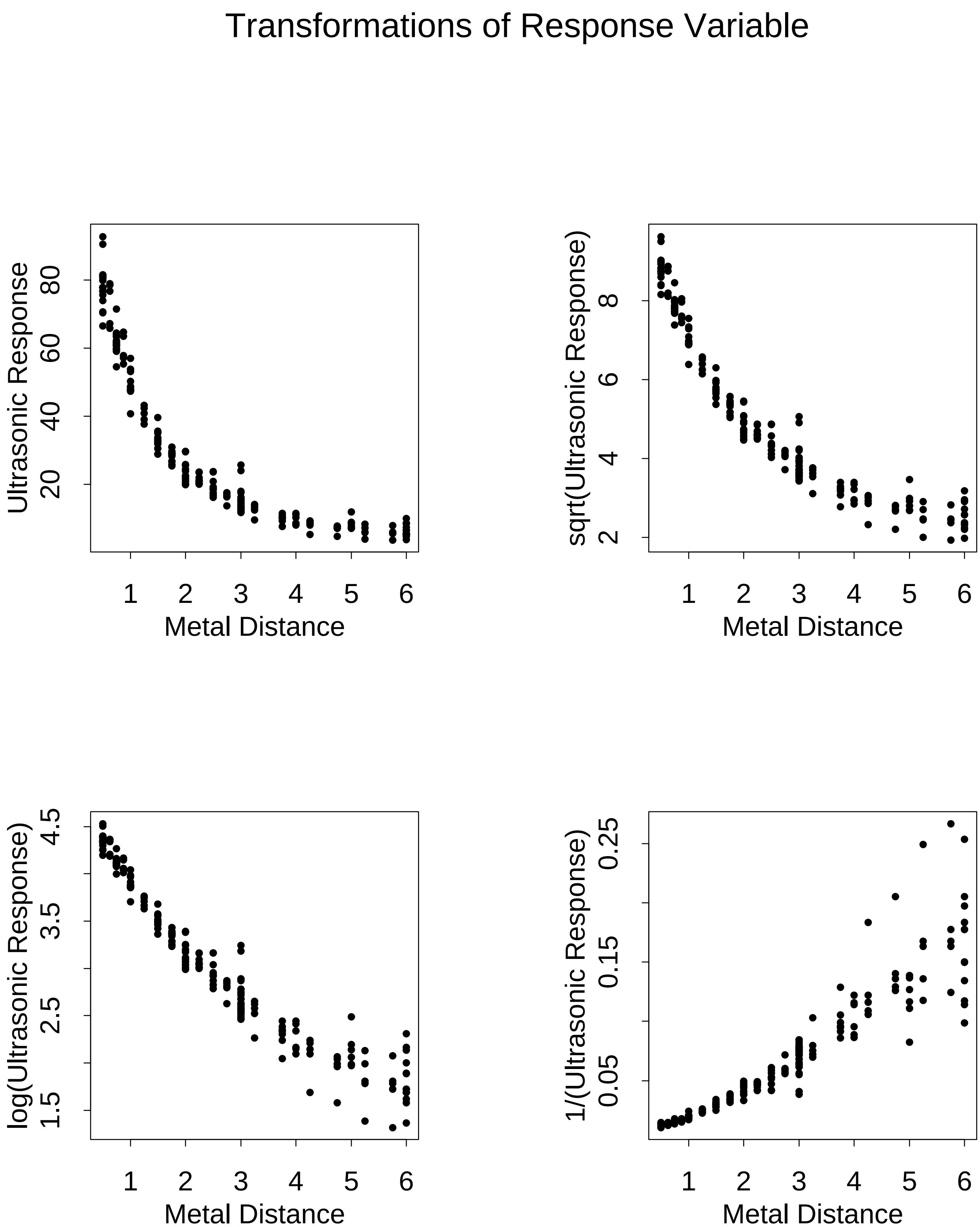

Data with Unweighted Line, WLS Fit,
and Fit Using Transformed Variables
Unweighted Line
Weighted Fit
Transformed Fit
Field Defect Size
Lab Defect Size
20
40
60
80
20
40
60
80


Ultrasonic Calibration Data
with Unweighted Nonlinear Fit
Ultrasonic Response
20
40
60
80
1
2
3
4
5
6
Metal Distance

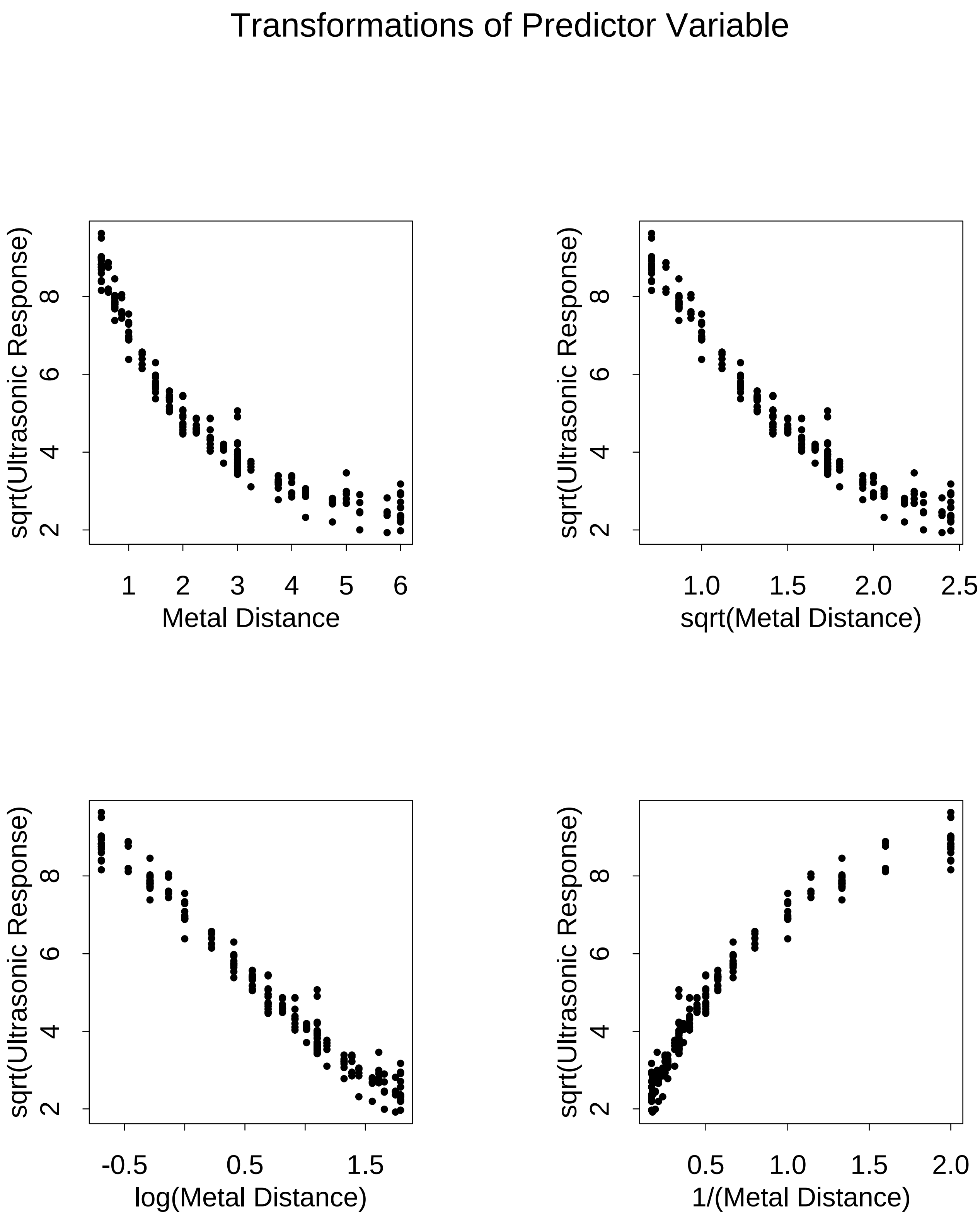

Ultrasonic Calibration Data Residuals - Unweighted Fit
Residuals
10
5
0
-5
-10
1
2
3
4
5
6
Metal Distance

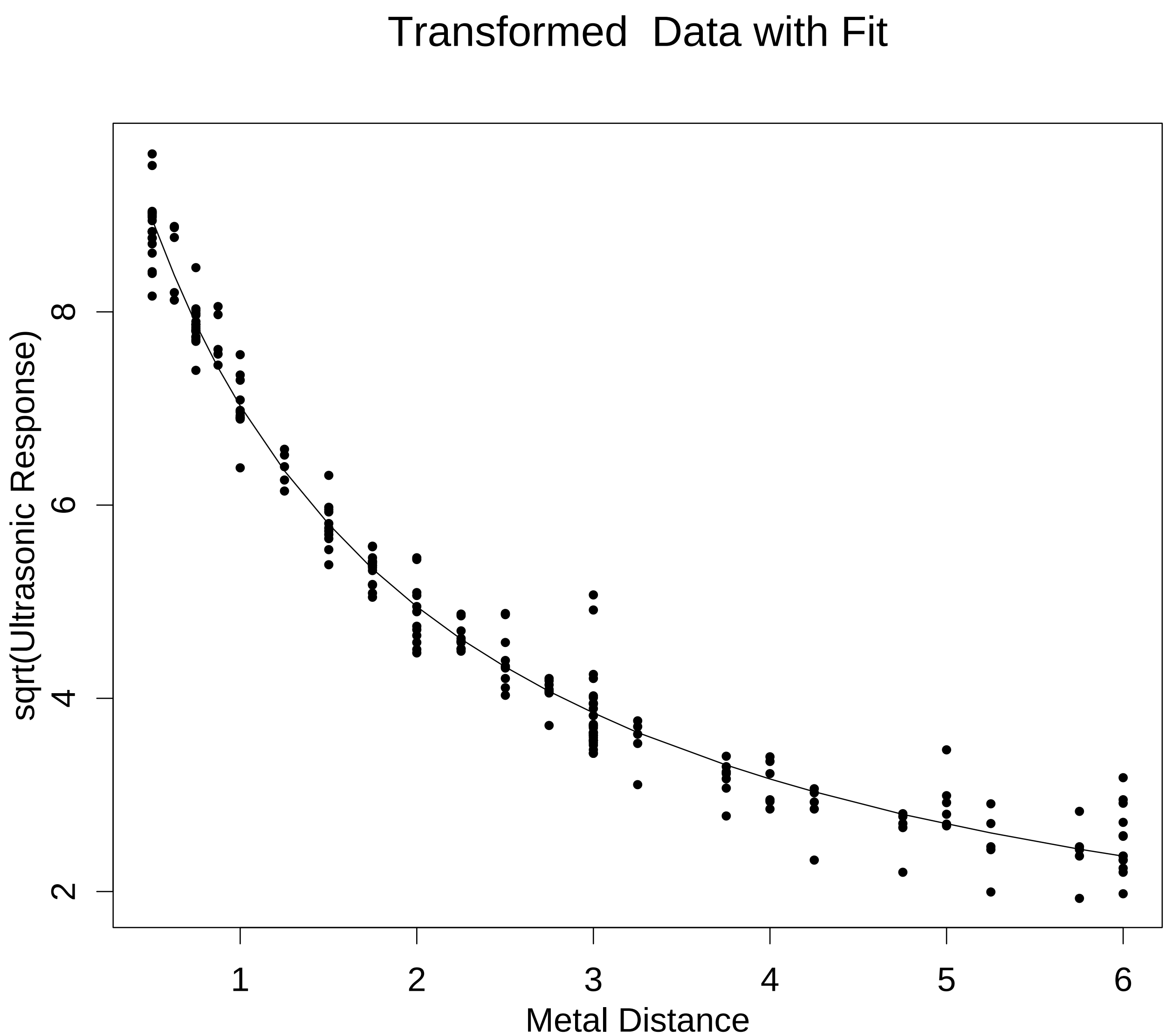
Ultrasonic Calibration Data
Output from Unweighted Exp/Linear Fit

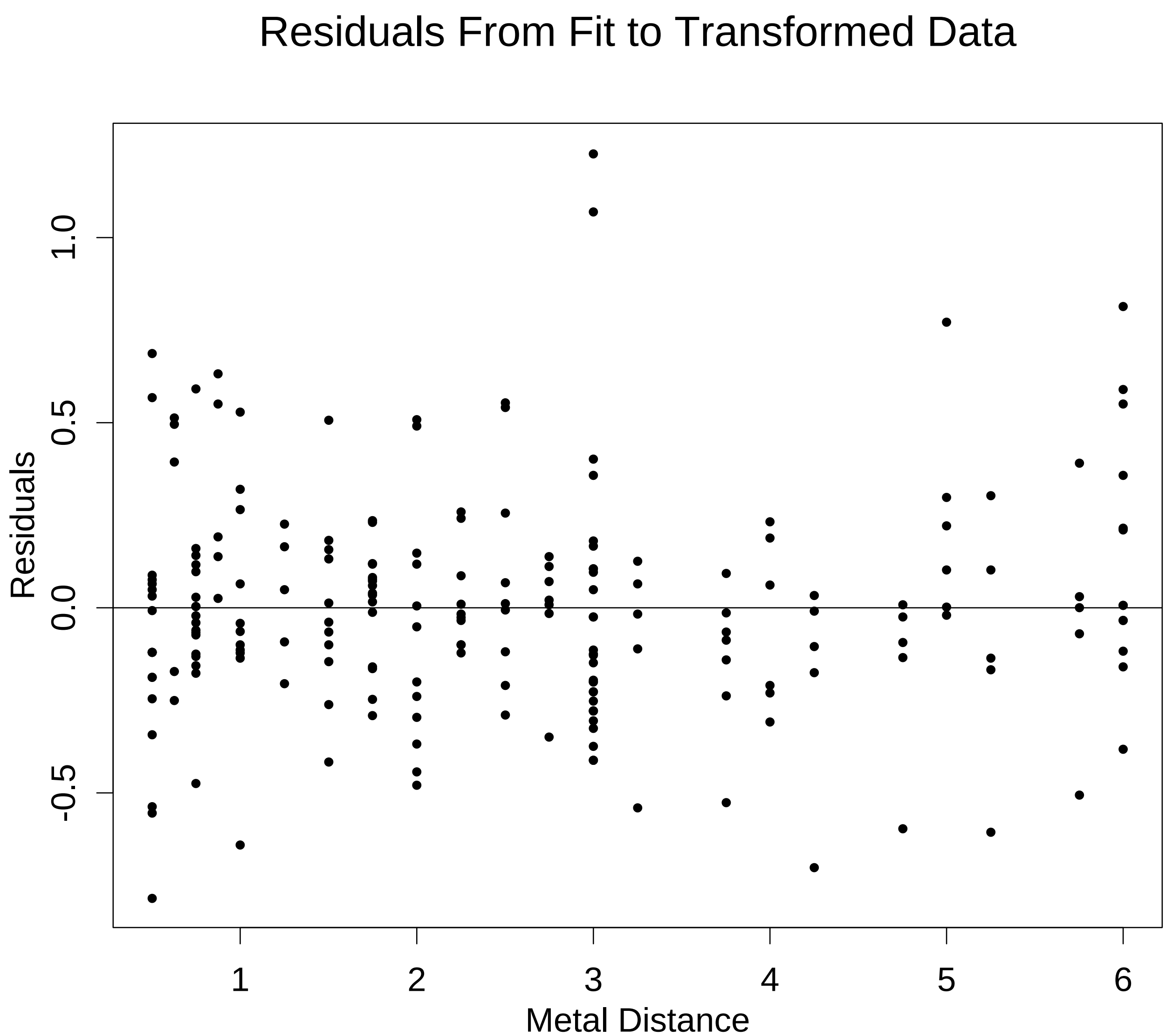
```
Parameters:

        Value  Std. Error  t value
b1 0.19027400 0.021938300  8.67312
b2 0.00613137 0.000345001 17.77200
b3 0.01053100 0.000792818 13.28300

Residual standard error: 3.36167 on 211 degrees of freedom

Correlation of Parameter Estimates:

       b1     b2
b2  0.839
b3 -0.950 -0.949
```

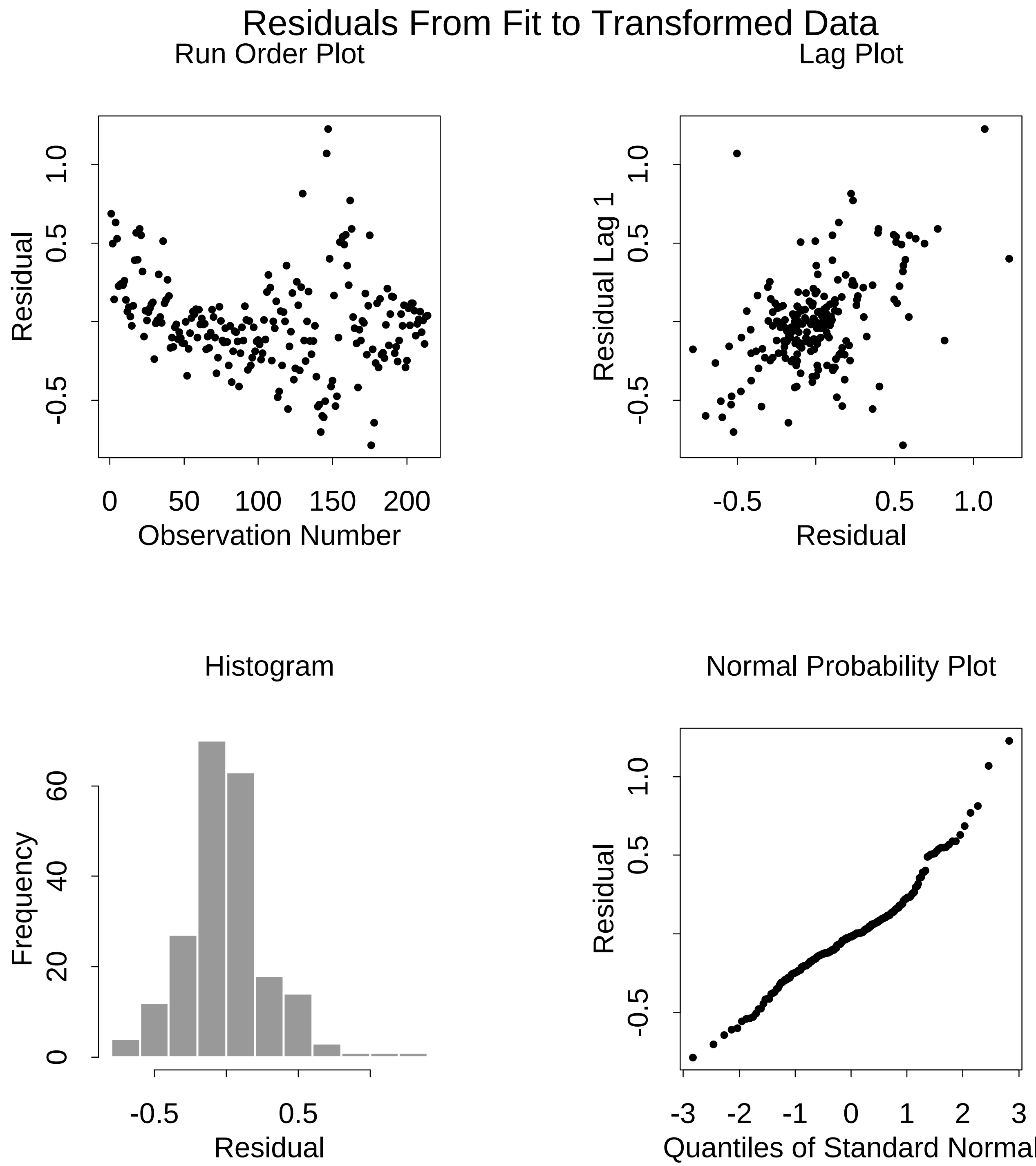


Transformations of Response Variable
Ultrasonic Response
sqrt(Ultrasonic Response)
log(Ultrasonic Response)
1/(Ultrasonic Response)
Metal Distance
Metal Distance
Metal Distance
Metal Distance


Transformations of Predictor Variable
sqrt(Ultrasonic Response)
2 4 6 8
1 2 3 4 5 6
Metal Distance
sqrt(Ultrasonic Response)
2 4 6 8
1.0 1.5 2.0 2.5
sqrt(Metal Distance)
sqrt(Ultrasonic Response)
2 4 6 8
-0.5 0.5 1.5
log(Metal Distance)
sqrt(Ultrasonic Response)
2 4 6 8
0.5 1.0 1.5 2.0
1/(Metal Distance)

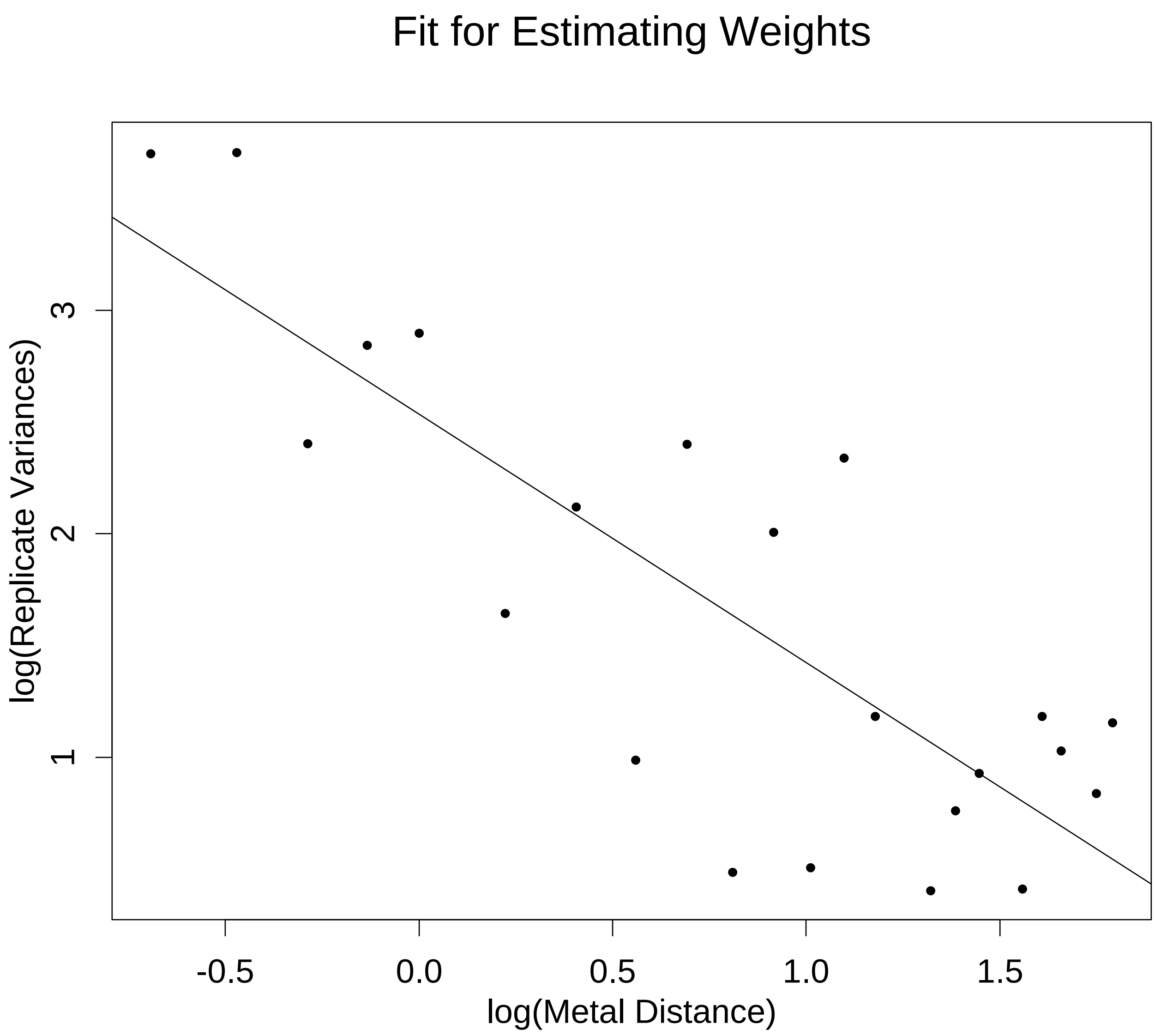

Transformed Data with Fit
sqrt(Ultrasonic Response)
Metal Distance
2
4
6
8
1
2
3
4
5
6

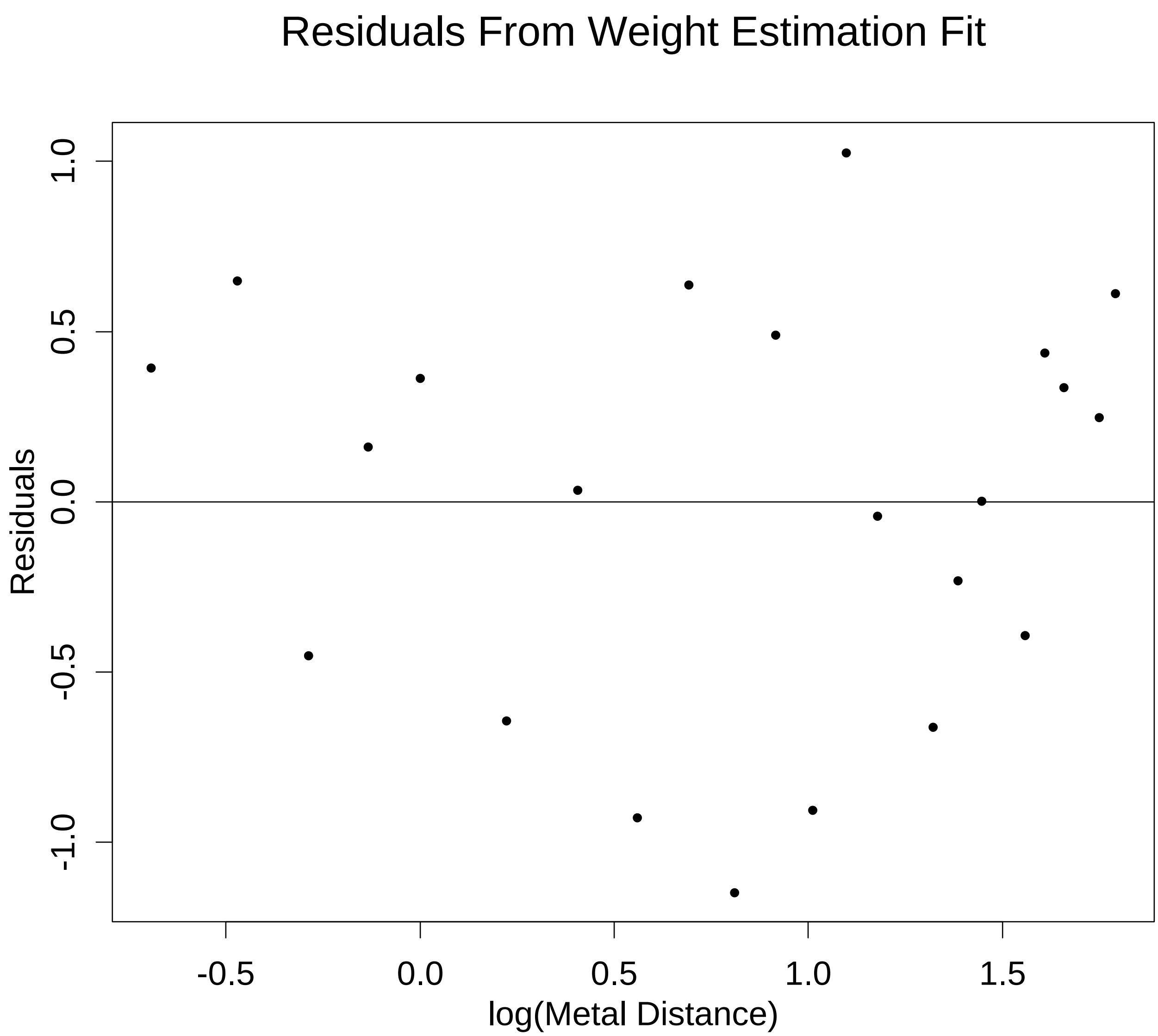

Residuals From Fit to Transformed Data
Residuals
1.0
0.5
0.0
-0.5
1
2
3
4
5
6
Metal Distance


Residuals From Fit to Transformed Data
Run Order Plot
Residual
-0.5
0.5
1.0
0
50
100
150
200
Observation Number
Lag Plot
Residual Lag 1
-0.5
0.5
1.0
-0.5
0.5
1.0
Residual
Histogram
Frequency
0
20
40
60
-0.5
0.5
Residual
Normal Probability Plot
Residual
-0.5
0.5
1.0
-3
-2
-1
0
1
2
3
Quantiles of Standard Normal

Ultrasonic Calibration Data
Output from Fit to Transformed Data

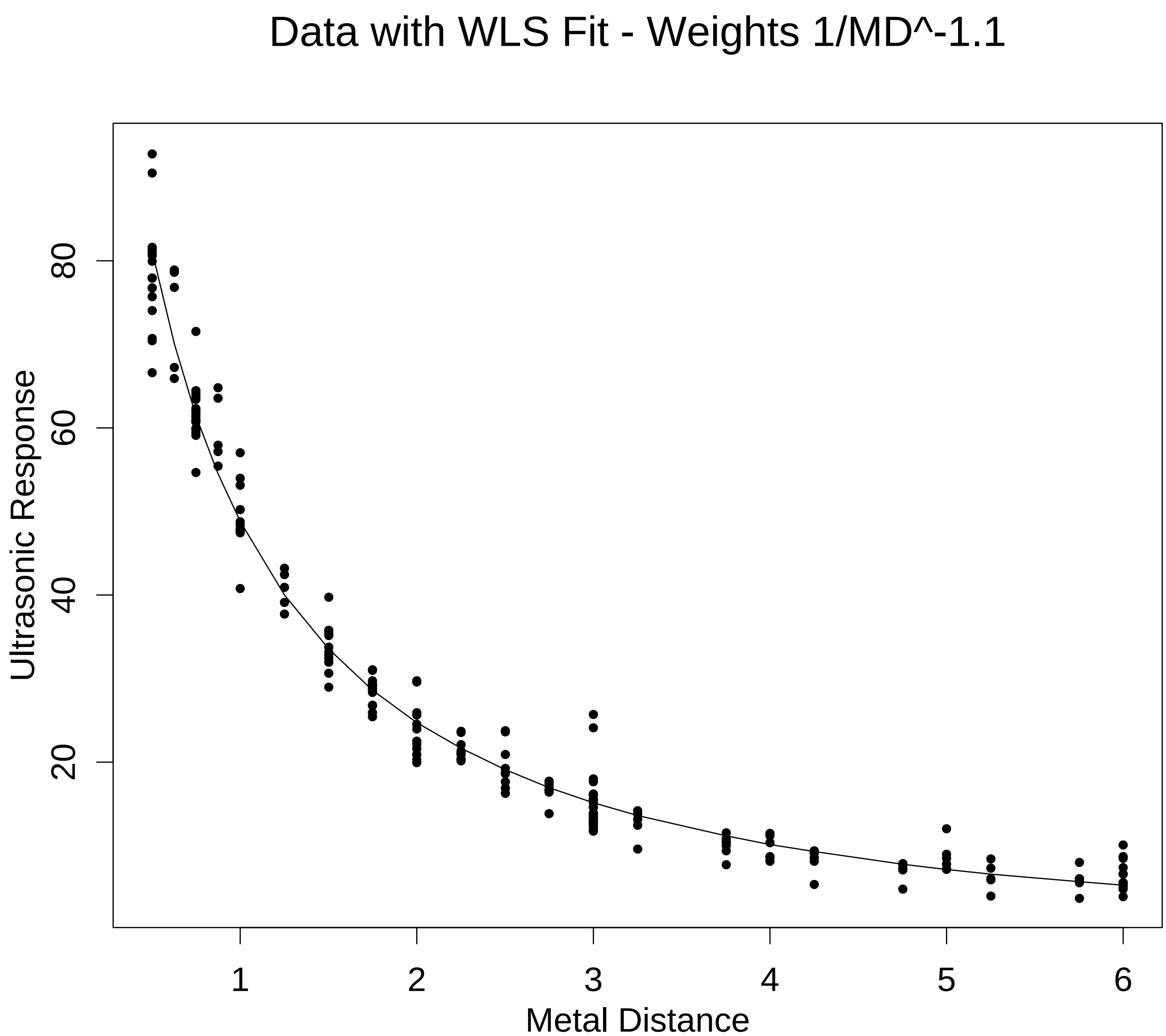
```
Parameters:

        Value Std. Error   t value
b1 -0.0154274 0.00861101  -1.79159
b2  0.0806725 0.00150574  53.57670
b3  0.0638570 0.00288001  22.17250

Residual standard error: 0.29715 on 211 degrees of freedom

Correlation of Parameter Estimates:

       b1     b2
b2  0.793
b3 -0.960 -0.899
```

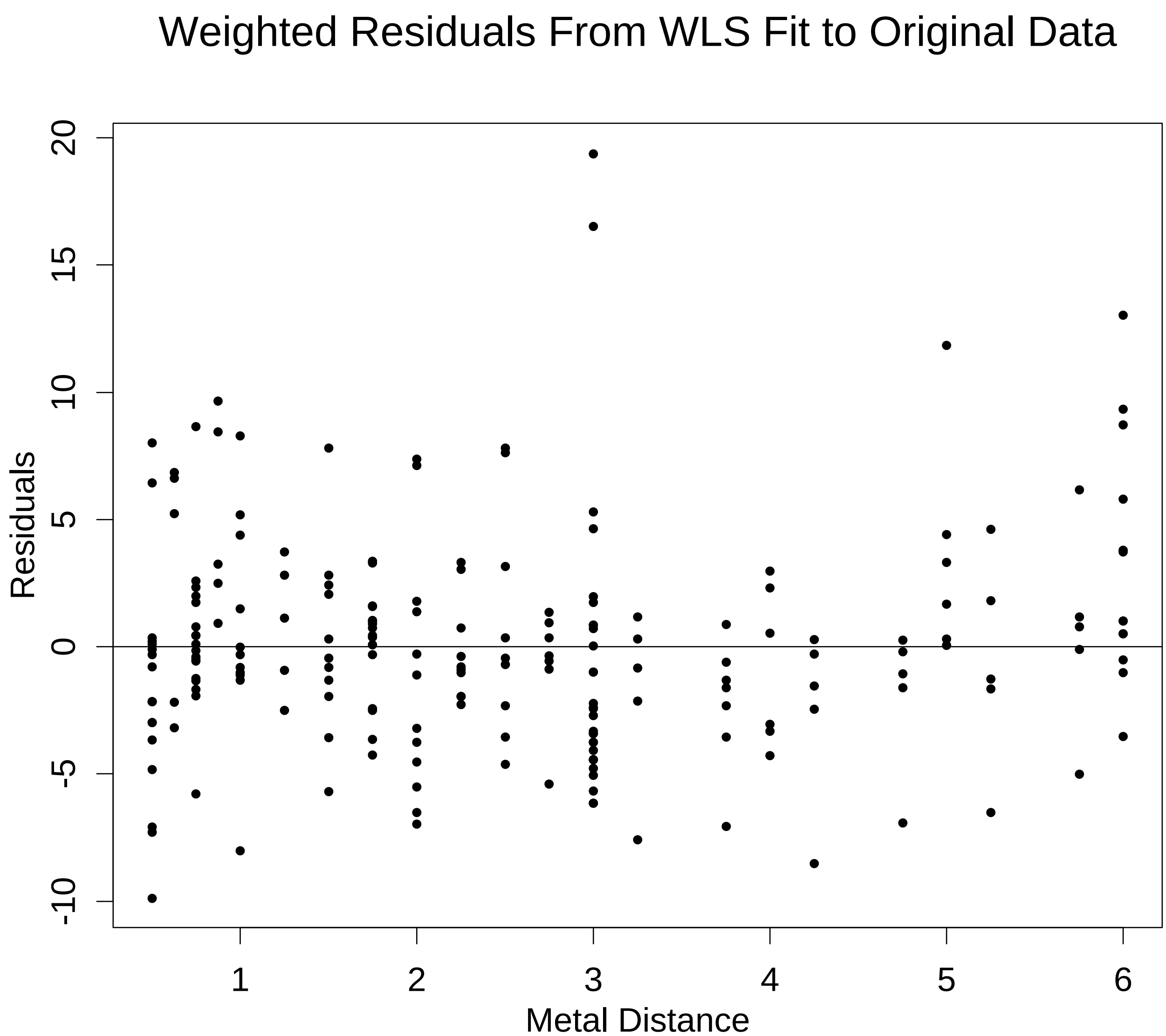

Fit for Estimating Weights
log(Replicate Variances)
1
2
3
-0.5
0.0
0.5
1.0
1.5
log(Metal Distance)

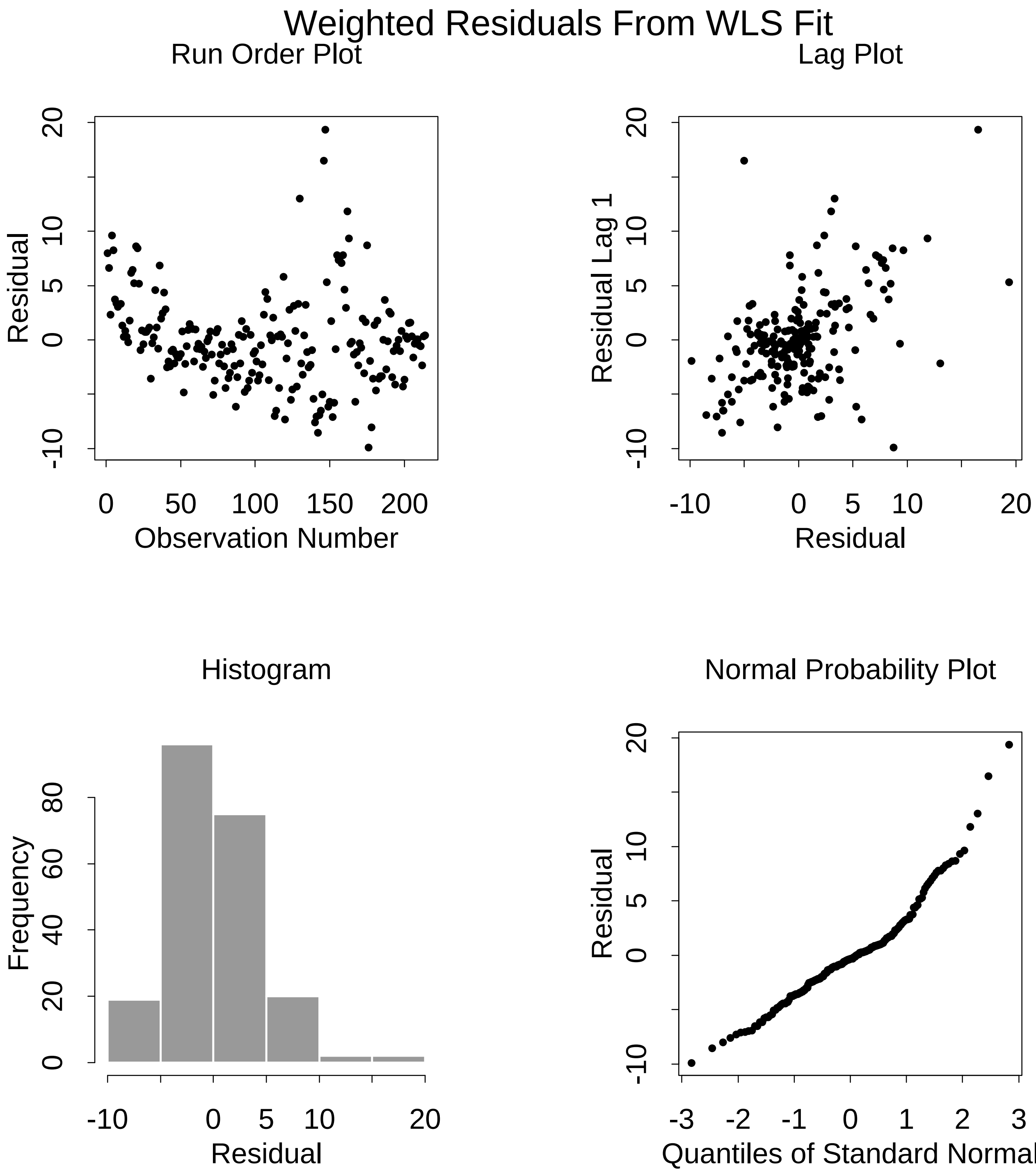

Residuals From Weight Estimation Fit
Residuals
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5
log(Metal Distance)

Ultrasonic Calibration Data Output from Weight Estimation Fit

```
N = 22

Residual Standard Error = 0.6099457

Multiple R-Square = 0.6712342

F-statistic = 40.83357 on 1 and 20 df, p-value = 3.10602e-06

               coef   std.err     t.stat      p.value
Intercept  2.536866 0.1919360  13.217253 2.421219e-11
        X -1.112763 0.1741382  -6.390115 3.106020e-06
```


Data with WLS Fit - Weights 1/MD^-1.1
Ultrasonic Response
20
40
60
80
1
2
3
4
5
6
Metal Distance

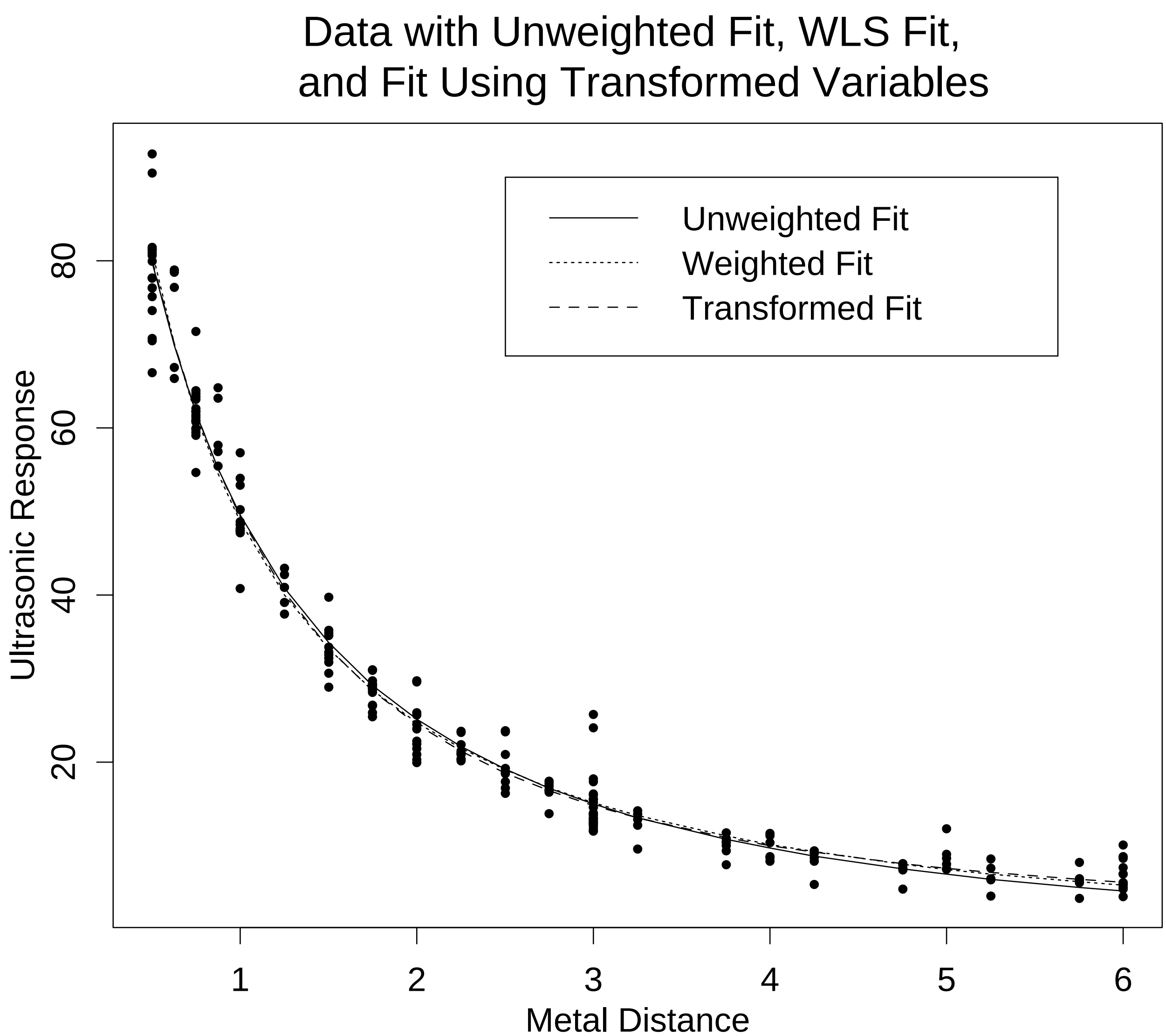

Weighted Residuals From WLS Fit to Original Data
Residuals
Metal Distance
20
15
10
5
0
-5
-10
1
2
3
4
5
6



Weighted Residuals From WLS Fit
Run Order Plot
Residual
Observation Number
Lag Plot
Residual Lag 1
Residual
Histogram
Frequency
Residual
Normal Probability Plot
Residual
Quantiles of Standard Normal

Ultrasonic Calibration Data
Output from Weighted Fit

```
LEAST SQUARES NON-LINEAR FIT
SAMPLE SIZE N =      214
MODEL--Y =EXP(-B1*X)/(B2+B3*X)
REPLICATION CASE
REPLICATION STANDARD DEVIATION =     0.3281762600D+01
REPLICATION DEGREES OF FREEDOM =         192
NUMBER OF DISTINCT SUBSETS     =          22


FINAL PARAMETER
      ESTIMATES     (APPROX. ST. DEV.)    T VALUE
1  B1  0.143378       (0.1476E-01)          9.7
2  B2  0.518479E-02   (0.4191E-03)          12.
3  B3  0.125719E-01   (0.7508E-03)          17.

RESIDUAL    STANDARD DEVIATION =         4.2653684616
RESIDUAL    DEGREES OF FREEDOM =         211
REPLICATION STANDARD DEVIATION =         3.2817625999
REPLICATION DEGREES OF FREEDOM =         192
LACK OF FIT F RATIO =  8.6545 = THE 100.0000% POINT OF THE
F DISTRIBUTION WITH     19 AND    192 DEGREES OF FREEDOM
```


Data with Unweighted Fit, WLS Fit,
and Fit Using Transformed Variables
Unweighted Fit
Weighted Fit
Transformed Fit
Ultrasonic Response
20
40
60
80
1
2
3
4
5
6
Metal Distance

Outliers

In a broad sense, outliers are points that do not follow the general pattern or structure in the data.

If present in the data, outliers can unduly affect the model-building process and invalidate predictions or calibrations.

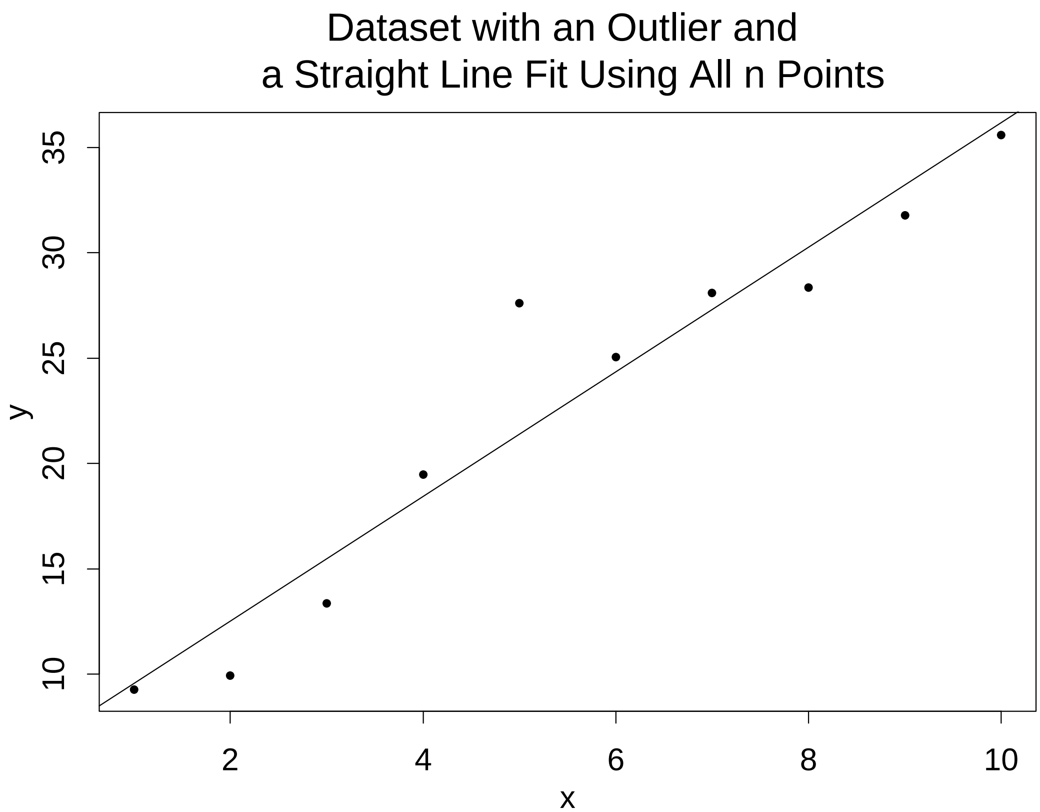
More Specifically . . .

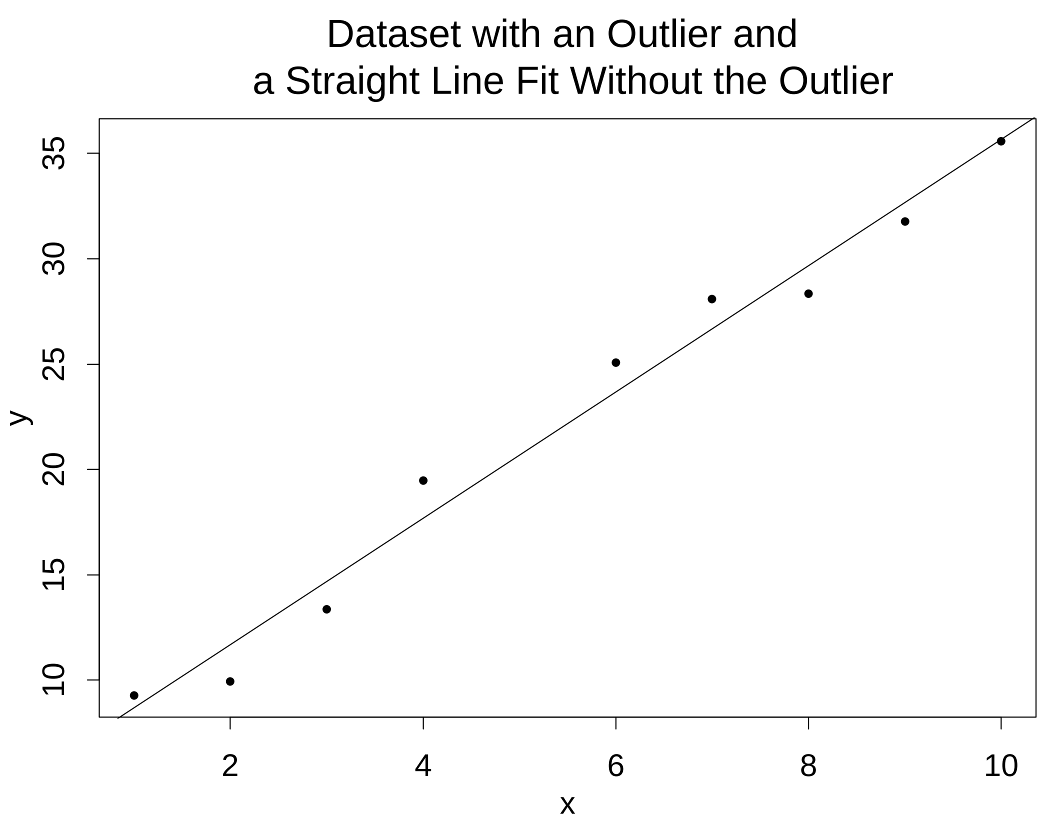
outlier - an observed value of the response variable that appears to be unusually large or small relative to its apparent population.

high leverage point - a point from the observed set of predictor variables that is unusual in size or structure relative to the other predictor variable points.

influential observation - an observation that has a large effect on the model derived from the data.

Note: Not all outliers or high leverage points are influential observations.


Dataset with an Outlier and
a Straight Line Fit Using All n Points
y
x
10
15
20
25
30
35
2
4
6
8
10

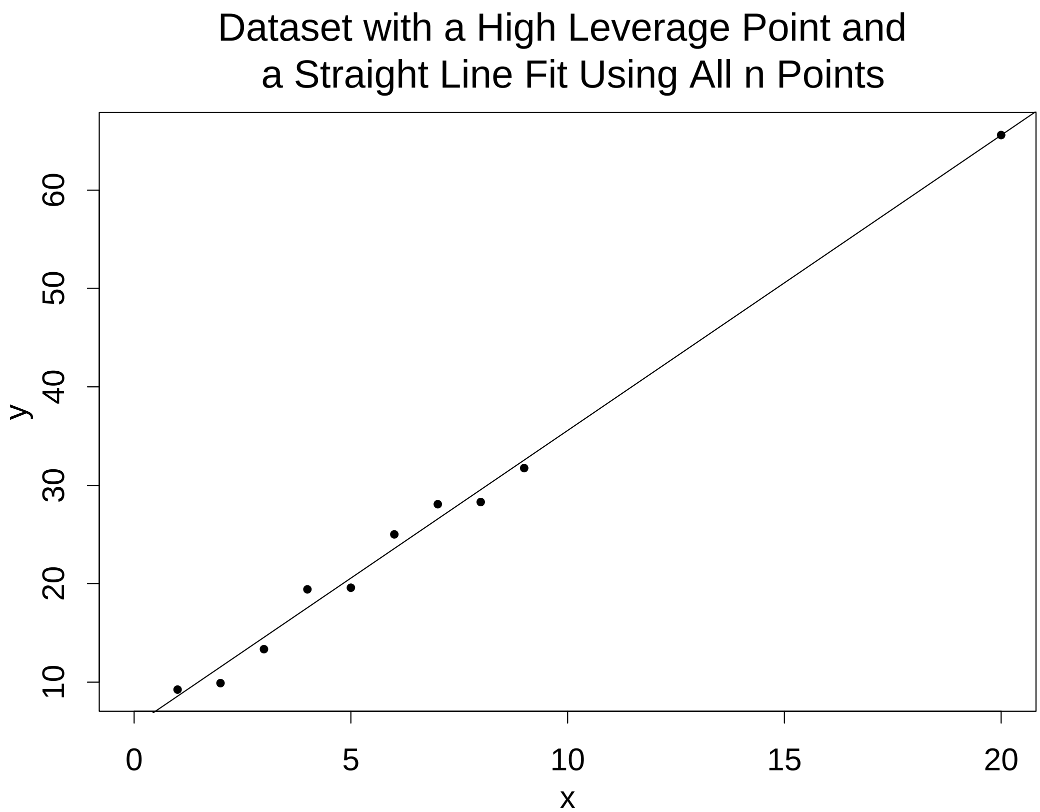

Dataset with an Outlier and
a Straight Line Fit Without the Outlier
y
x
10
15
20
25
30
35
2
4
6
8
10

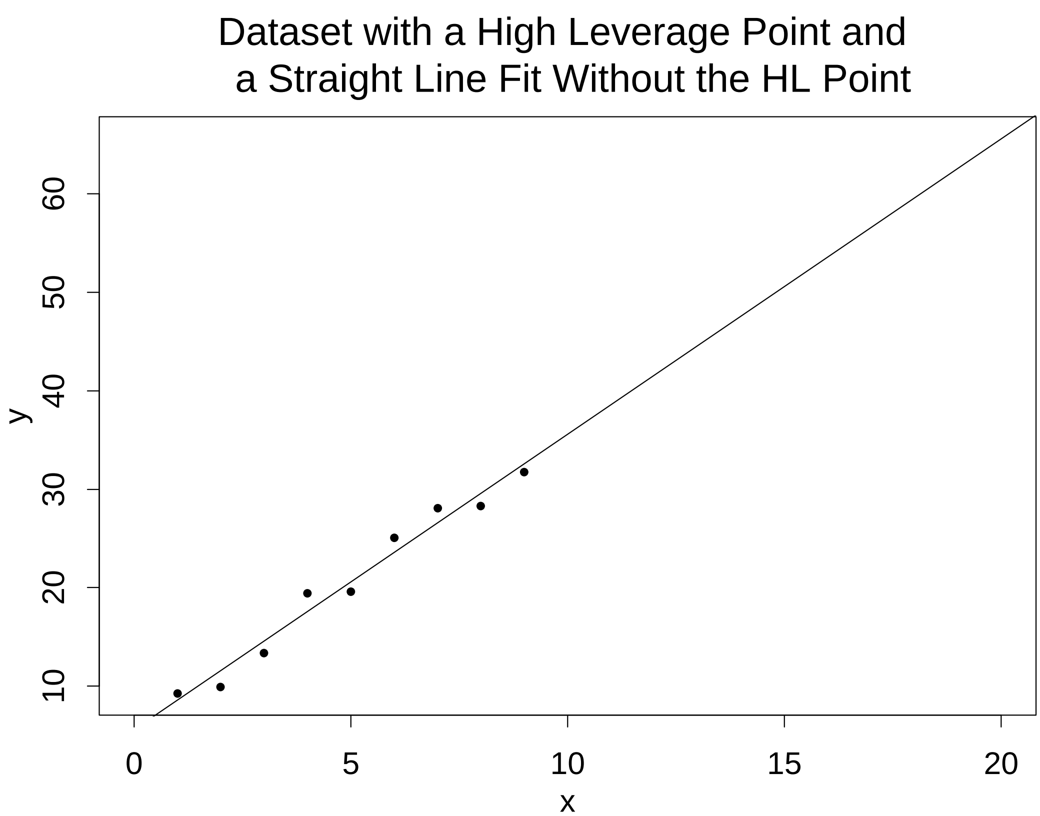
Regression Output for Data with Outlier

```
N = 10

Residual Standard Error = 2.6986

Multiple R-Square = 0.9252

F-statistic = 98.9308 on 1 and 8 df, p-value = 0

            coef std.err t.stat p.value
Intercept 6.6162  1.8435 3.5889  0.0071
    x     2.9552  0.2971 9.9464  0.0000
```

Regression Output for Data with Outlier Omitted

```
N = 9

Residual Standard Error = 1.4547

Multiple R-Square = 0.9803

F-statistic = 349.0913 on 1 and 7 df, p-value = 0

            coef std.err  t.stat p.value
Intercept 5.6882  1.0146  5.6063   8e-04
    x     2.9974  0.1604 18.6840   0e+00
```


Dataset with a High Leverage Point and
a Straight Line Fit Using All n Points
y
10
20
30
40
50
60
x
0
5
10
15
20

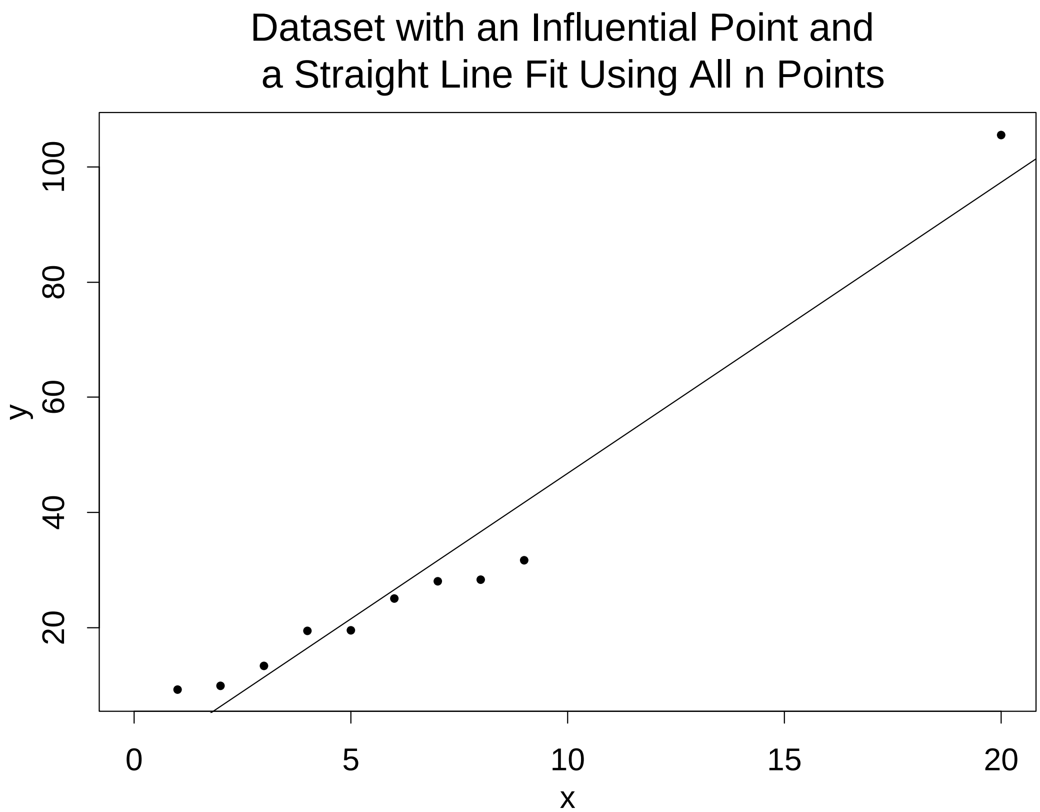

Dataset with a High Leverage Point and
a Straight Line Fit Without the HL Point
y
10
20
30
40
50
60
x
0
5
10
15
20

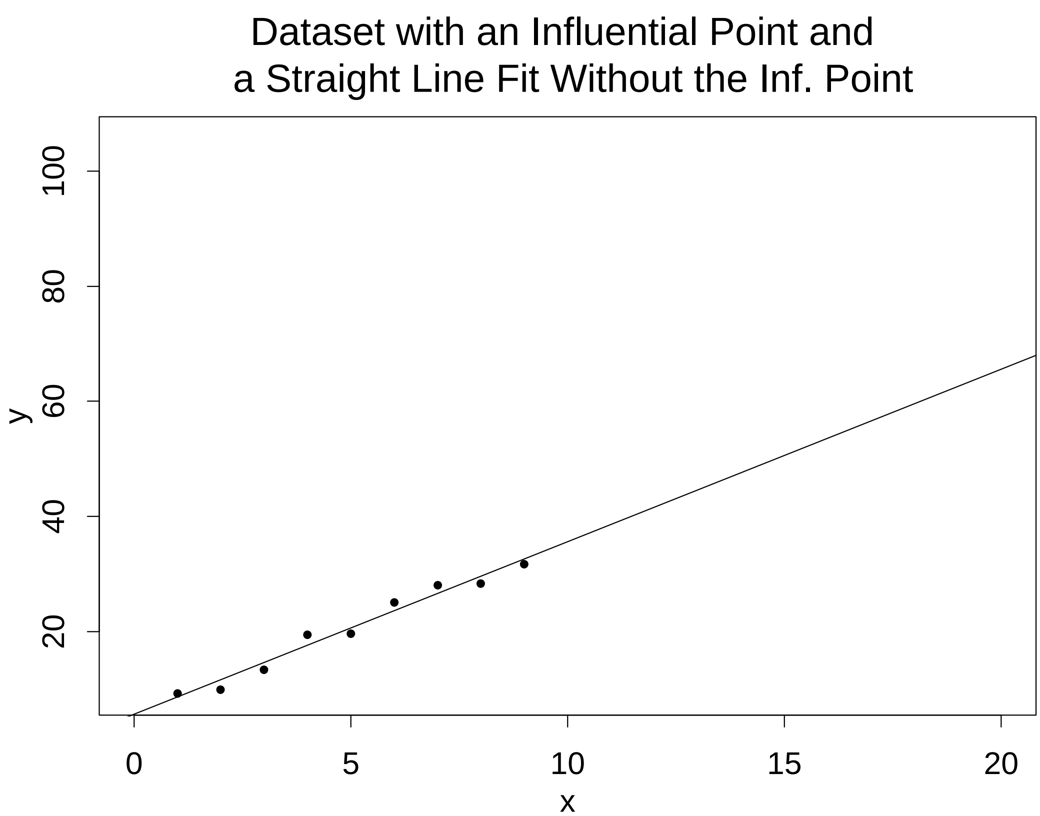
Regression Output for Data with Leverage Point

```
N = 10

Residual Standard Error = 1.4046

Multiple R-Square = 0.9934

F-statistic = 1199.086 on 1 and 8 df, p-value = 0

            coef std.err  t.stat p.value
Intercept 5.5567  0.7175  7.7444   1e-04
    x     3.0020  0.0867 34.6278   0e+00
```

Regression Output for Data with LP Omitted

```
N = 9

Residual Standard Error = 1.5016

Multiple R-Square = 0.9717

F-statistic = 240.0134 on 1 and 7 df, p-value = 0

            coef std.err  t.stat p.value
Intercept 5.5512  1.0909  5.0888  0.0014
    x     3.0032  0.1939 15.4924  0.0000
```


Dataset with an Influential Point and
a Straight Line Fit Using All n Points
y
100
80
60
40
20
0
5
10
15
20
x


Dataset with an Influential Point and
a Straight Line Fit Without the Inf. Point
y
100
80
60
40
20
0
5
10
15
20
x

Regression Output for Data with Influential Point

```
N = 10

Residual Standard Error = 6.5626

Multiple R-Square = 0.9512

F-statistic = 156.0045 on 1 and 8 df, p-value = 0

             coef std.err   t.stat p.value
Intercept -3.8147  3.3524  -1.1379  0.2881
    x      5.0591  0.4050  12.4902  0.0000
```

Regression Output for Data with IP Omitted

```
N = 9

Residual Standard Error = 1.5016

Multiple R-Square = 0.9717

F-statistic = 240.0134 on 1 and 7 df, p-value = 0

            coef std.err  t.stat p.value
Intercept 5.5512  1.0909  5.0888  0.0014
    x     3.0032  0.1939 15.4924  0.0000
```

Why Outliers Affect LS Regression

As we've seen before, the least squares estimates are found by minimizing the quantity Q to find the $\hat{\beta}$'s.

$$Q = \sum_{i=1}^{n}[y_i - f(x_{1i}, x_{2i}, \ldots, x_{ki}; \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)]^2$$

However, when a few deviations are much larger than the rest (in absolute value), and all of the deviations are squared and added up, the few large deviations tend to dominate the total.

As a result, the values of the $\hat{\beta}$'s that minimize of the least squares criterion are often the values that eliminate large deviations from the model.

Why Outliers Affect LS Regression

The LSS estimates for the straight line regression function $y = \beta_1 + \beta_2 x$ show in detail how both outliers and high leverage points affect the parameter estimates.

$$\hat{\beta}_2 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$\hat{\beta}_2$ is a linear combination of the deviations of the data points, y_i, from their sample mean, $\bar{y}$, 'weighted' proportional to the deviation of x_i from $\bar{x}$. Large values of either $(x_i - \bar{x})$ or $(y_i - \bar{y})$ will impact the estimates of both β_1 and β_2.

The effects of leverage points and outliers are analogous in more complicated models.

General Steps for Handling Outliers

The basic steps for dealing with outliers in regression analysis are:

1. identify the outlying and influential points
2. determine the reason each of these points was observed, if possible
3. eliminate any points that can be shown to be irrelevant to the analysis
4. reanalyze the data, both with and without any remaining influential points

Steps 1 and 4 in this general procedure are statistical in nature, while steps 2 and 3 require knowledge of specific experimental details and a thorough understanding of the underlying science.

Outliers Detection Methods

Fortunately, outliers can usually be found using standard model validation procedures. However, in complicated, multivariate data, they can be sometimes be difficult to pick out.

Use of more specialized outlier detection procedures in linear regression problems can help ensure that all outliers are found, and can lead to a deeper understanding of the data and model.

The use and interpretation of the specialized outlier detection methods is harder in nonlinear least squares, though logically possible. As a result, residual plots are the primary means of identifying influential points in nonlinear problems.

Two Classes of Specialized Outlier Detection Methods

There are basically two classes of outlier detection methods:

1. methods that rely on the usual LS model fitting criteria, but may (effectively) delete observations from the fit, one-at-a-time, to help determine their influence (LS Methods), and

2. methods in which the model fitting criterion is changed to reduce the effect of outliers on model estimation (Robust Methods)

Advantages and Disadvantages of LS Methods

Advantages:

1. use the same software as for regular regression analysis
2. work reasonably well with small and large data sets
3. provide a relatively large amount of information about unusual data points
4. are part of a relatively well-developed area of statistical research
5. are computationally efficient

Disadvantages:

1. do not work well with clusters of unusual points
2. require use of many different diagnostic statistics

Advantages and Disadvantages of Robust Methods

Advantages:

1. work well with arbitrary numbers and clusters of unusual points
2. identify influential points in one step

Disadvantages:

1. require specialized software
2. require medium-sized to large data sets
3. may not identify non-influential outlying points
4. are part of a relatively new area of statistical research
5. can be computationally difficult

LS Methods for Outlier Detection

Four of the main LS statistics for outlier detection are:

leverages - which are used to identify high leverage points (unusual predictor variable values), regardless of their influence,

studentized deleted residuals - which are used to identify outliers (unusual response variable values) which may or may not be influential,

Cook's distances - which are used to identify points that influence one or more of the estimated parameters, and

DFFITS - which are used to identify points which influence one or more of the predicted values.

Leverage Values (aka Hat Values)

Leverage values are the diagonal values of the Hat matrix, H, which is given by:

$$H = X(X^T X)^{-1} X^T$$

where X is the matrix with n rows and $p = k + 1$ columns, the first of which is a column of 1's for the intercept term, and each of the other columns given by the values of one of the k variables in the model,

$$X = [1|x_1|x_2|\ldots|x_k]$$

For the straight line model

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

Leverages for more complicated models are conceptually analogous to the straight line case.

Leverage Values (aka Hat Values)

The sum of the leverage values is p, the number of parameters in the model, for all multivariable linear regression functions.

Therefore, when an individual leverage value is much larger than the mean leverage p/n, it indicates a high leverage point.

A good cut-off for leverage values is $2.5p/n$. Points with values larger than this are likely to be contributing to the fitted model more heavily than most.

High leverage doesn't necessarily indicate an influential point, however. A high leverage point with a y value that is right on target will produce a fit that is similar to one using only low leverage points.

Studentized Deleted Residuals

The i^{th} studentized deleted residual is conceptually obtained by:

1. deleting the i^{th} point from the data,
2. fitting the model with the other $n - 1$ observations,
3. computing the deleted residual, which is the difference of the i^{th} response and corresponding prediction from the fit made without the i^{th} response.
4. computing the residual standard deviation using the fit to the $n - 1$ other points, and
5. dividing the deleted residual by its standard deviation, which depends on the residual standard deviation with the i^{th} point deleted and the i^{th} leverage value computed using all of the data.

Studentized Deleted Residuals

The conceptual formula for the i^{th} studentized deleted residual is

$$T_i = \frac{y_i - \hat{y}_{(-i)}(x_{1i}, \ldots, x_{ki})}{s_{(-i)}/\sqrt{1 - h_{ii}}}$$

where $\hat{y}_{(-i)}(x_{1i}, \ldots, x_{ki})$ is the predicted value based on the fit without the i^{th} point, and $s_{(-i)}$ is the corresponding residual standard deviation.

The computational formula is:

$$T_i = e_i \sqrt{\frac{n - p - 1}{(n - p)s^2(1 - h_{ii}) - e_i^2}}$$

where the computations are done using all of the data.

Studentized Deleted Residuals

Technically, T_i has a t distribution with $n - p - 1$ degrees of freedom, and t distribution cut-off values can be used to determine if a particular point is an outlier.

Practically speaking, studentized deleted residuals greater than 2.5 in absolute value indicate that a point is likely to be an outlier.

Remember, however, that large studentized deleted residuals do not necessarily offer any information about the influence of the outlier on the fit of the model.

Cook's Distance

Conceptually, the i^{th} Cook's Distance measures the difference between the values of the parameters estimated using all n observations and the values of the parameters estimated without the i^{th} observation.

It provides an aggregate measure of the difference for all of the parameters simultaneously.

The computational formula for Cook's Distance is:

$$c_i = \frac{h_{ii}}{p}(\frac{e_i}{s(1-h_{ii})})^2$$

computed using all of the data.

Cook's Distance

Points with Cook's Distance values greater than the 50% F distribution cut-off with p and $n - p$ degrees of freedom are likely to influence the estimated parameters in the model.

That is, if a point with a large Cook's Distance is eliminated from the data set, the values of one or more of the parameters will change significantly.

Influential points should be studied carefully to determine if they are valid data points. If not, they should be excluded from the analysis. If there is no evidence that the influential points are invalid, further work will be required to obtain conclusive results from the data.

DFFITS

Conceptually, the i^{th} DFFITS value measures the difference between the i^{th} predicted value estimated using all n observations and estimated without the i^{th} observation.

The conceptual formula for DFFITS is:

$$\text{DFFITS}_i = \frac{\hat{y}_i(x_{1i}, \ldots, x_{ki}) - \hat{y}_{(-i)}(x_{1i}, \ldots, x_{ki})}{s_{(-i)}\sqrt{h_{ii}}}$$

The computational formula for DFFITS is:

$$\text{DFFITS}_i = T_i\sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

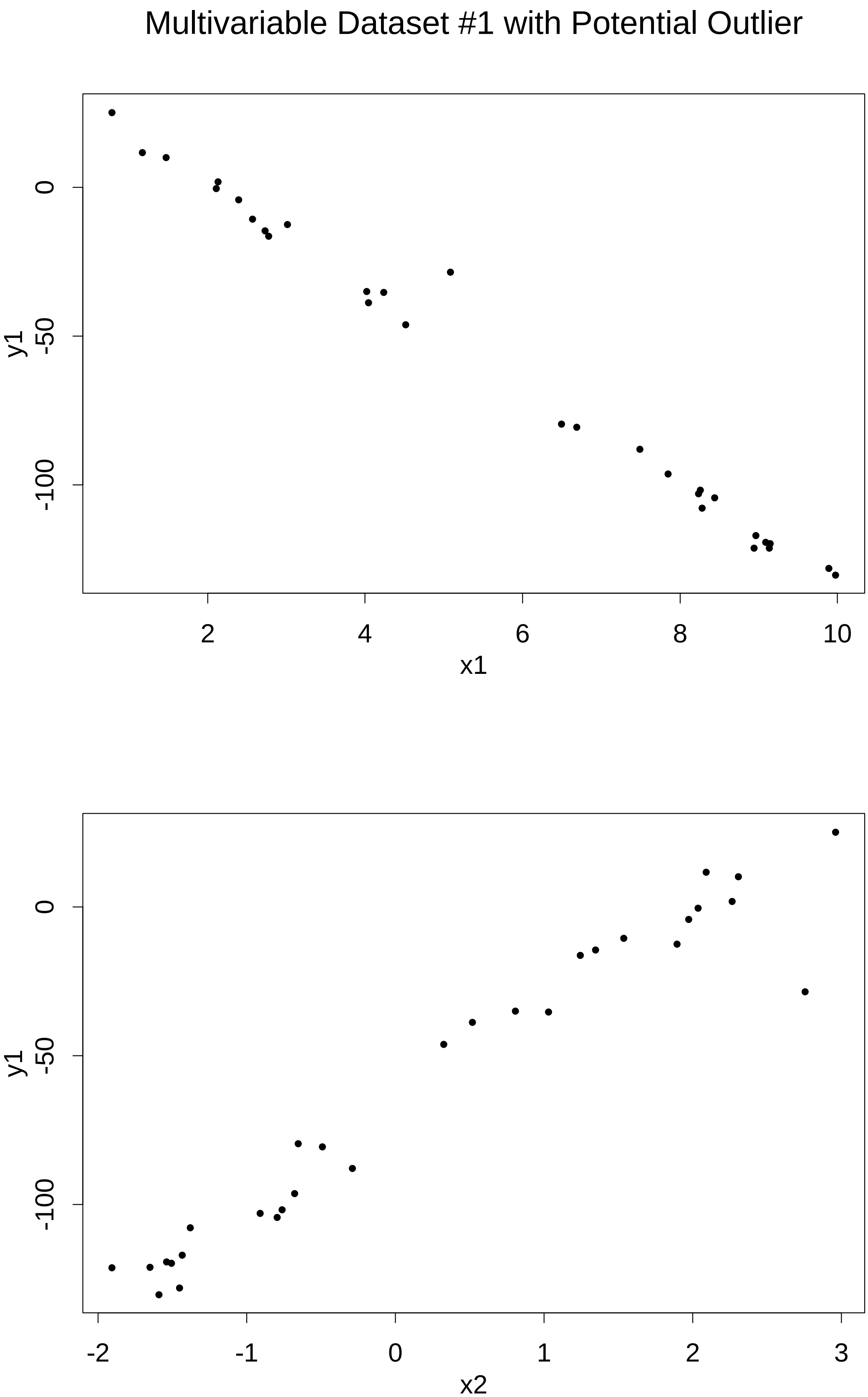
computed using all of the data.

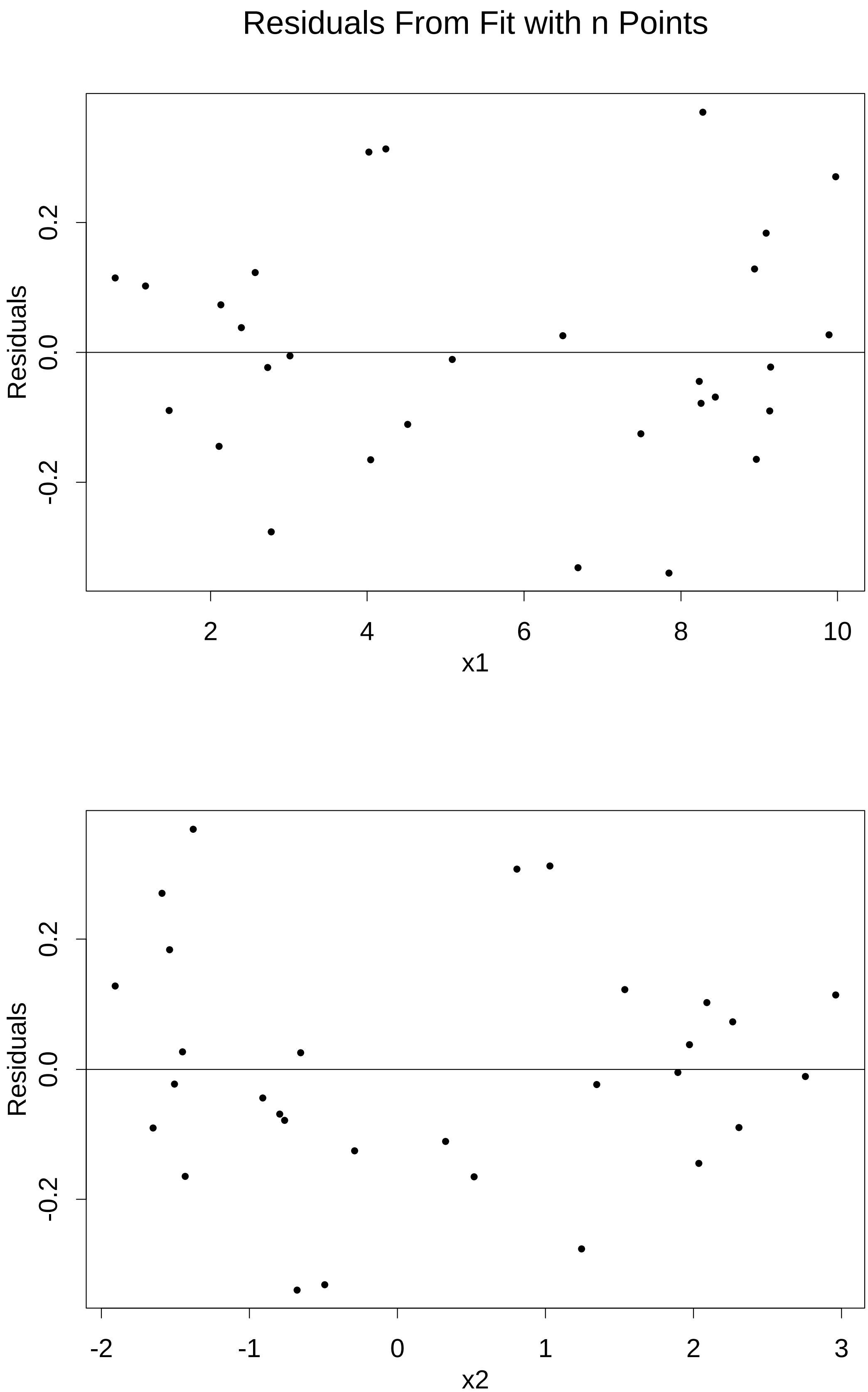
DFFITS

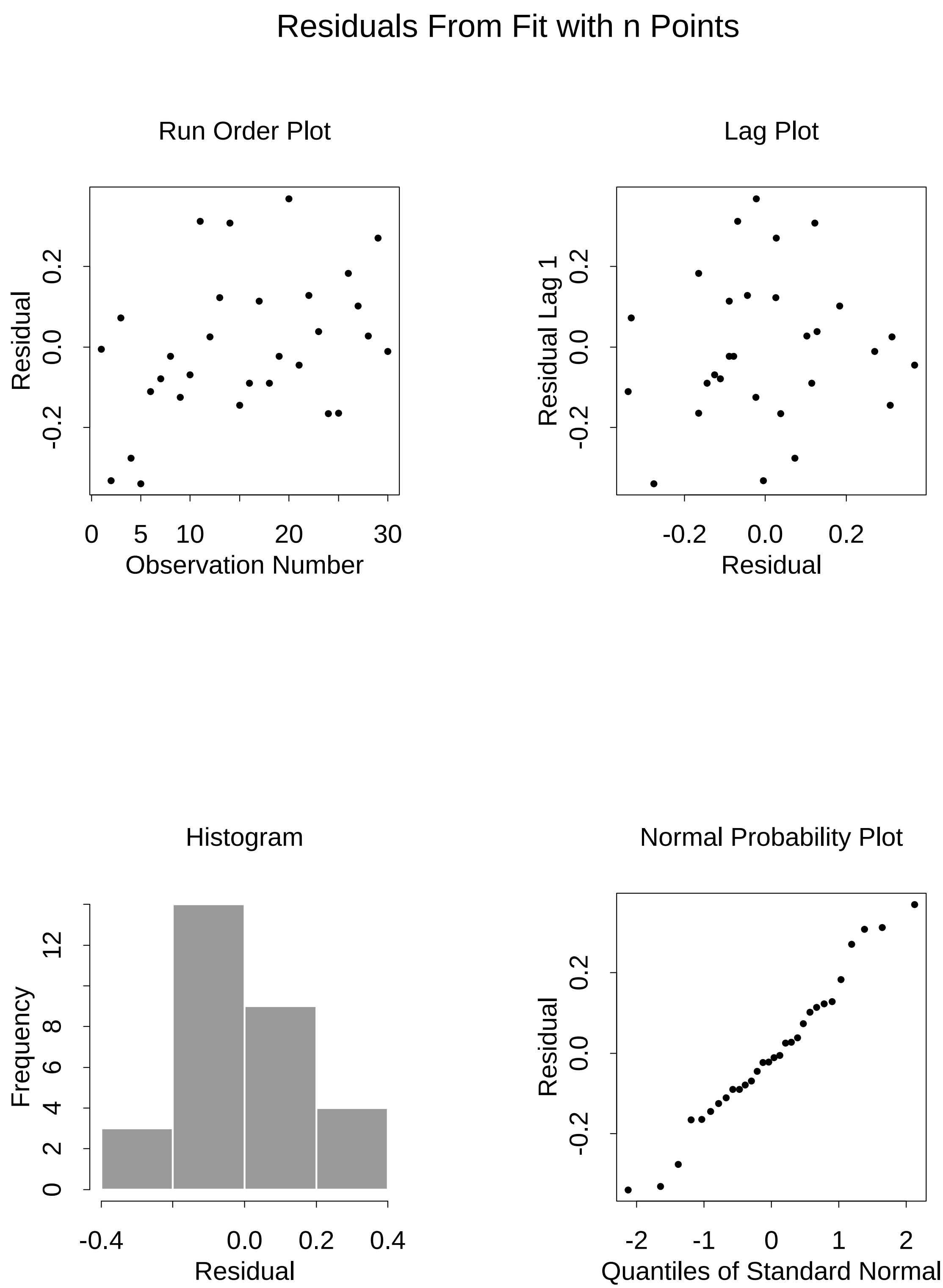
Points with DFFITS values greater than $2.5\sqrt{p/n}$ are likely to influence the predicted values from the model.

That is, if a point with a large DFFITS value is eliminated from the data set, the value of the one or more of the predicted values will change significantly.

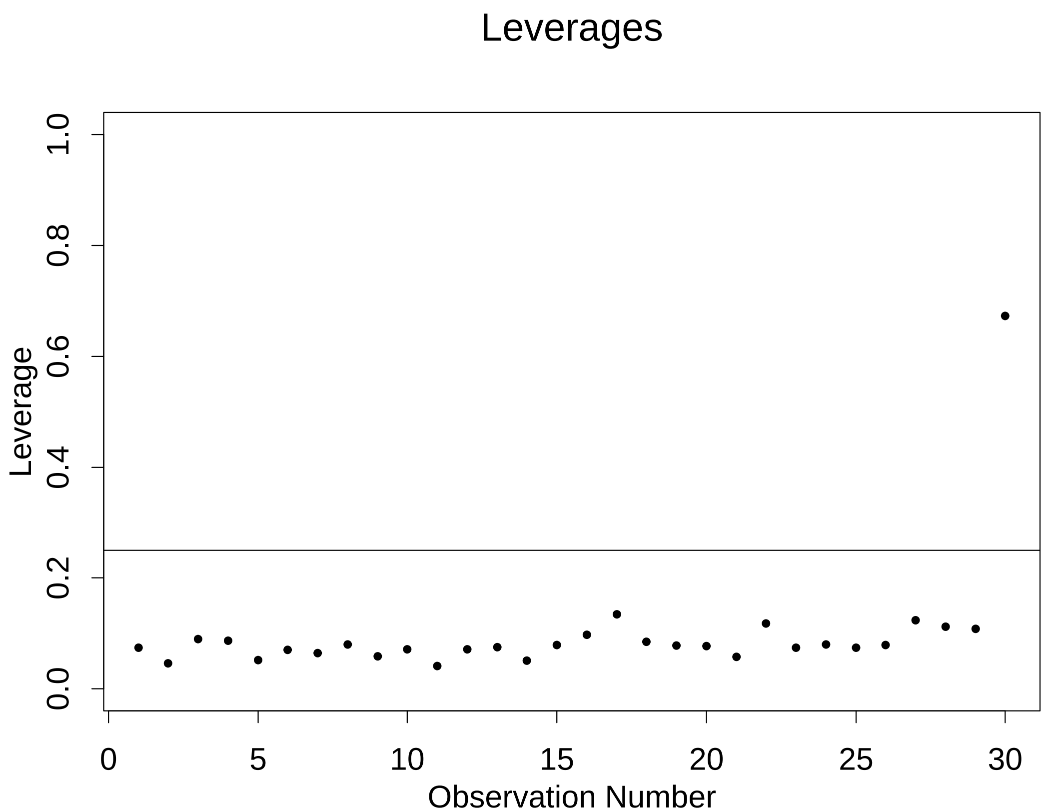
As with influential points identified using Cook's Distance, the validity of points picked out by DFFITS should be examined carefully. If assignable causes are found that indicate the point is not valid, it should be dropped from the analysis. If not, further work will required to obtain conclusive results.

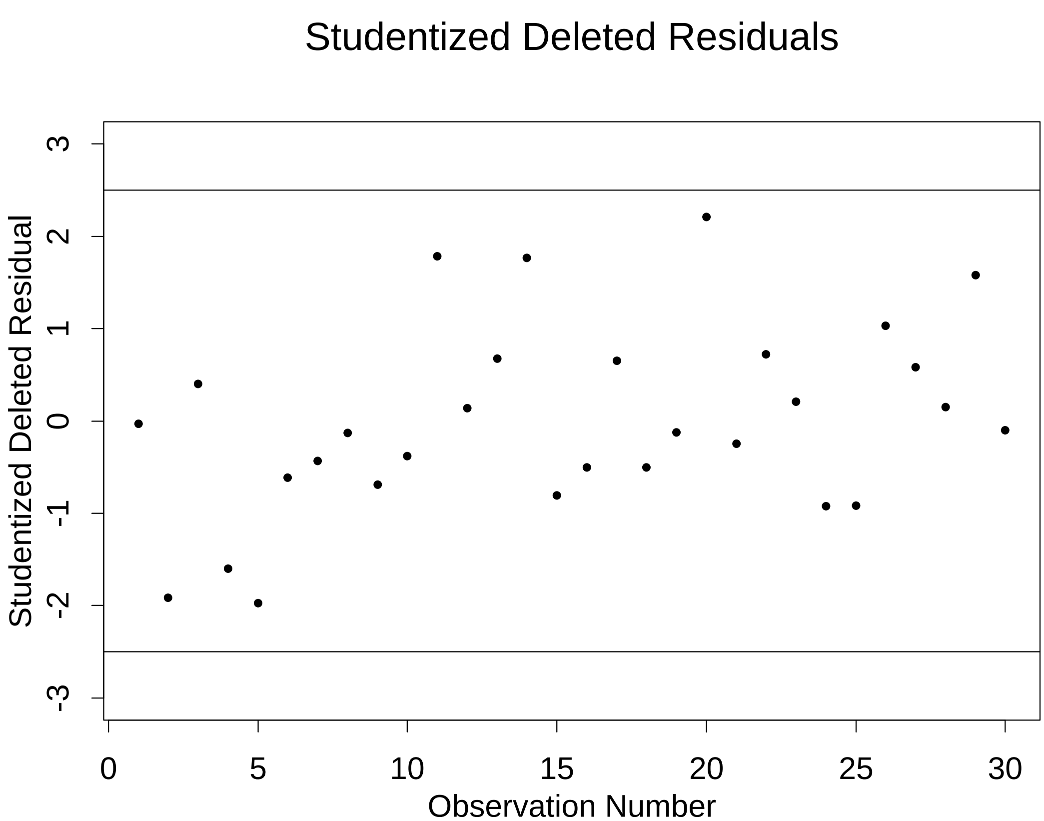

Multivariable Dataset #1 with Potential Outlier
y1
0
-50
-100
2
4
6
8
10
x1
y1
0
-50
-100
-2
-1
0
1
2
3
x2

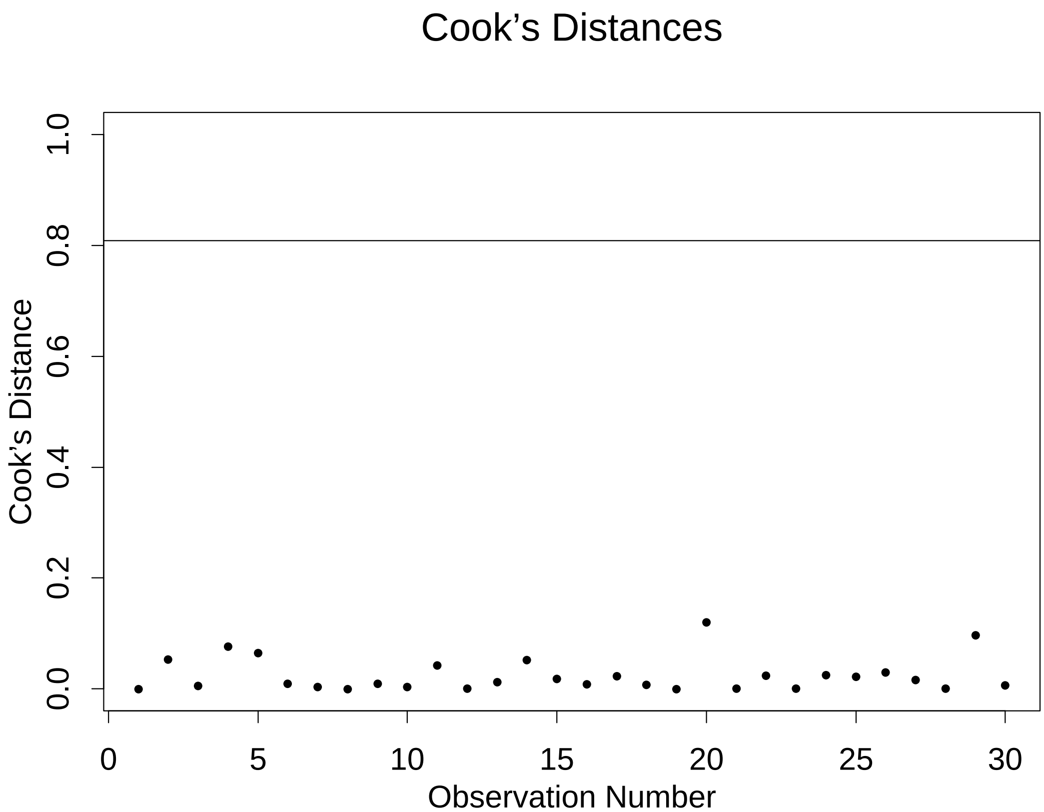

Residuals From Fit with n Points
Residuals
0.2
0.0
-0.2
2
4
6
8
10
x1
Residuals
0.2
0.0
-0.2
-2
-1
0
1
2
3
x2

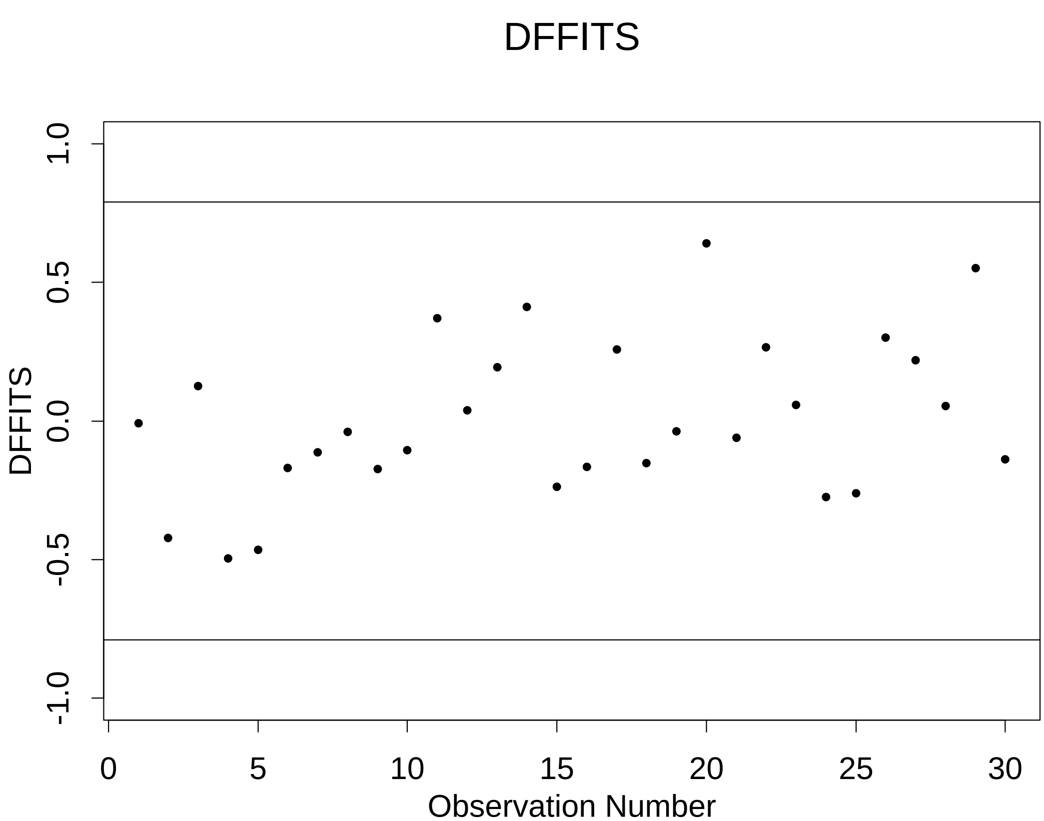


Residuals From Fit with n Points
Run Order Plot
Residual
-0.2
0.0
0.2
0
5
10
20
30
Observation Number
Lag Plot
Residual Lag 1
-0.2
0.0
0.2
-0.2
0.0
0.2
Residual
Histogram
Frequency
0
2
4
6
8
12
-0.4
0.0
0.2
0.4
Residual
Normal Probability Plot
Residual
-0.2
0.0
0.2
-2
-1
0
1
2
Quantiles of Standard Normal


Leverages
Leverage
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Observation Number


Studentized Deleted Residuals
Studentized Deleted Residual
-3
-2
-1
0
1
2
3
0
5
10
15
20
25
30
Observation Number


Cook's Distances
Cook's Distance
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Observation Number


DFFITS
DFFITS
-1.0
-0.5
0.0
0.5
1.0
0
5
10
15
20
25
30
Observation Number

Regression Output Using n Points

N = 30

Residual Standard Error = 0.1859

Multiple R-Square = 1

F-statistic = 1121069 on 2 and 27 df, p-value = 0

	coef	std.err	t.stat	p.value
Intercept	4.7251	0.2153	21.9480	0
x1	-11.9502	0.0345	-346.3767	0
x2	10.0313	0.0674	148.9384	0

Regression Output Using with the High Leverage Point Deleted

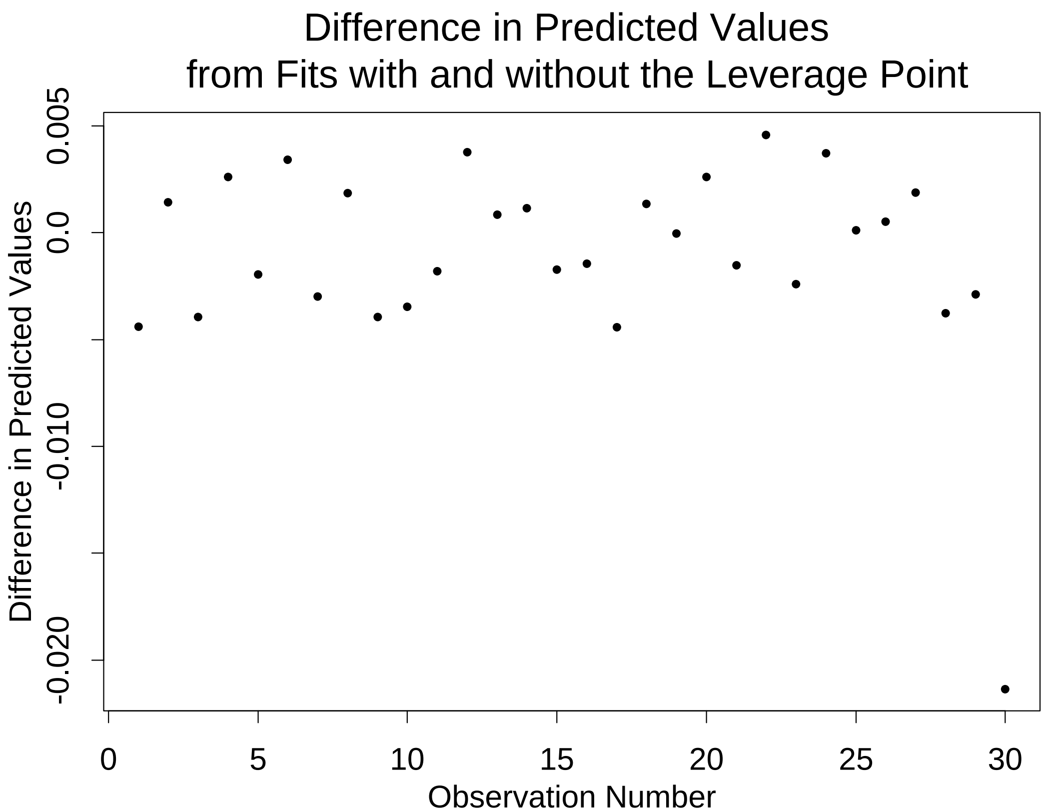
N = 29

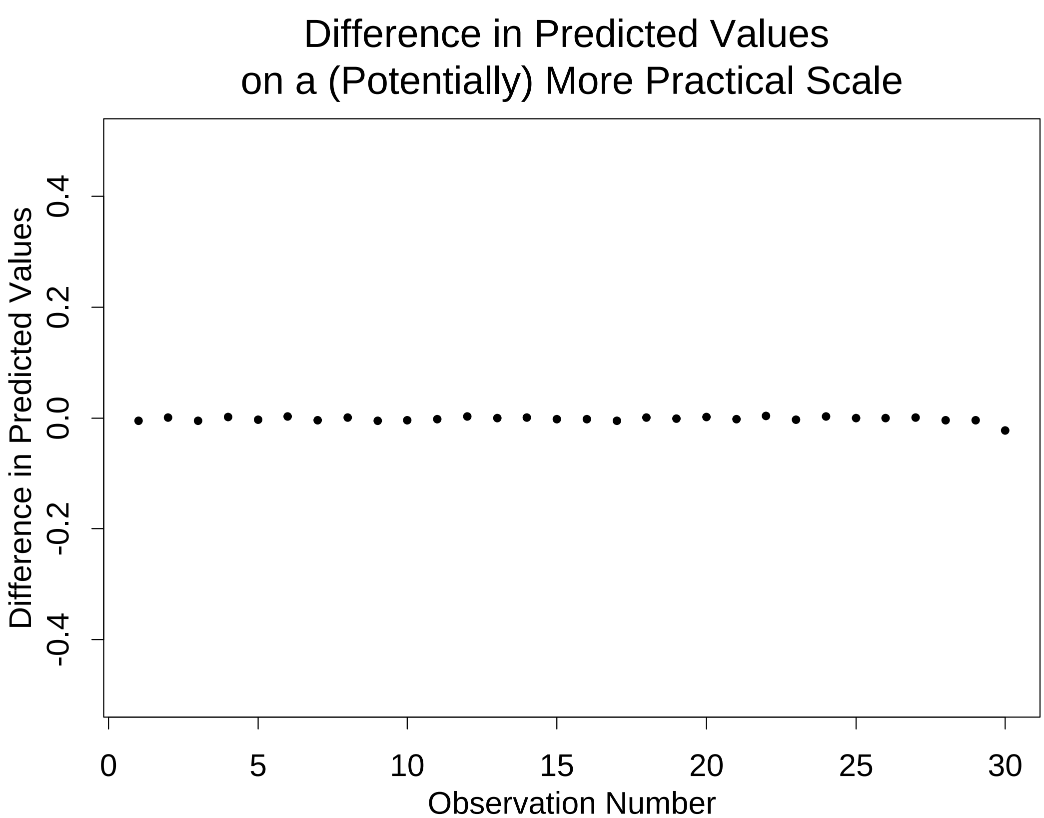
Residual Standard Error = 0.1895

Multiple R-Square = 1

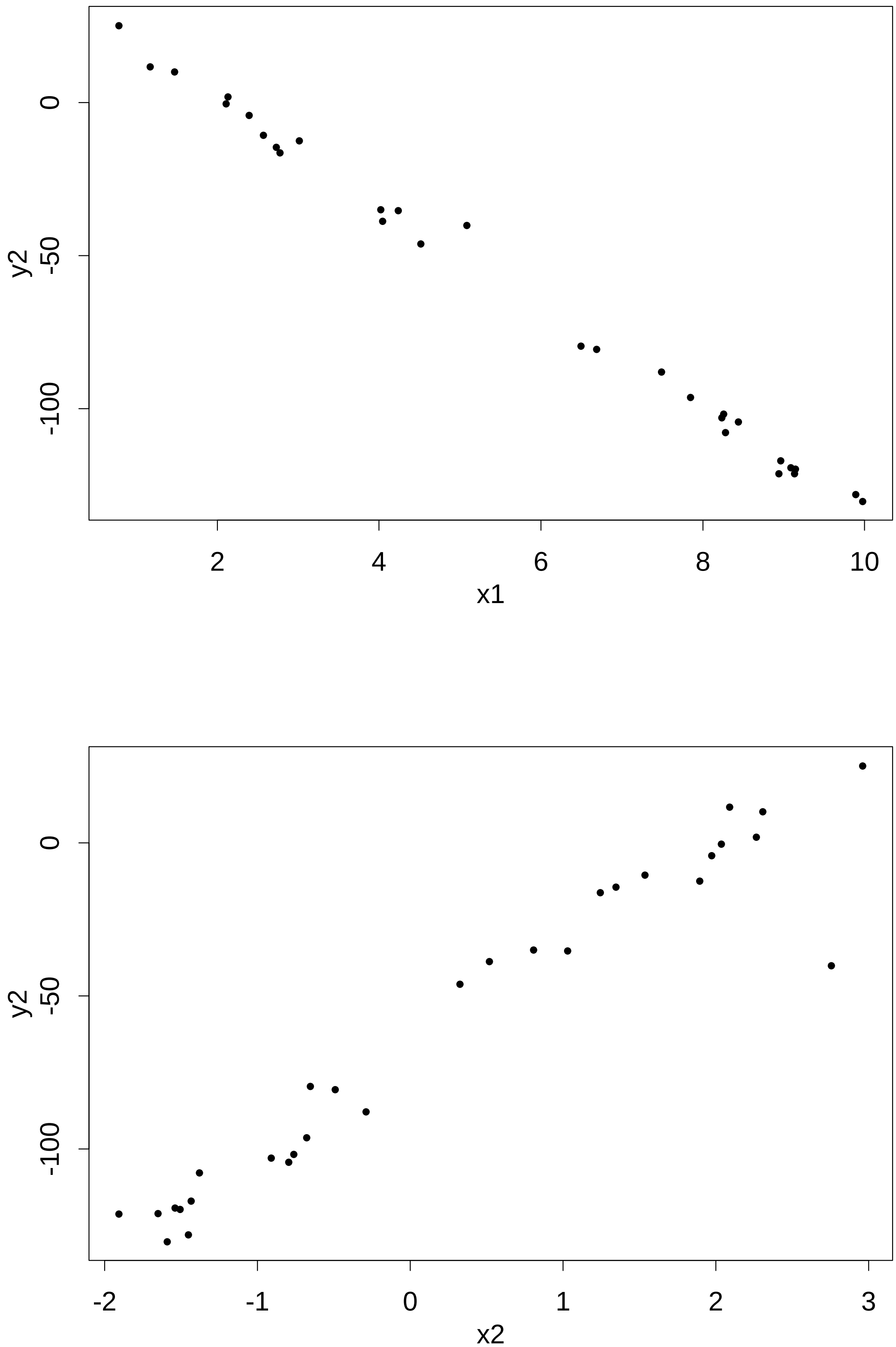
F-statistic = 1065271 on 2 and 26 df, p-value = 0

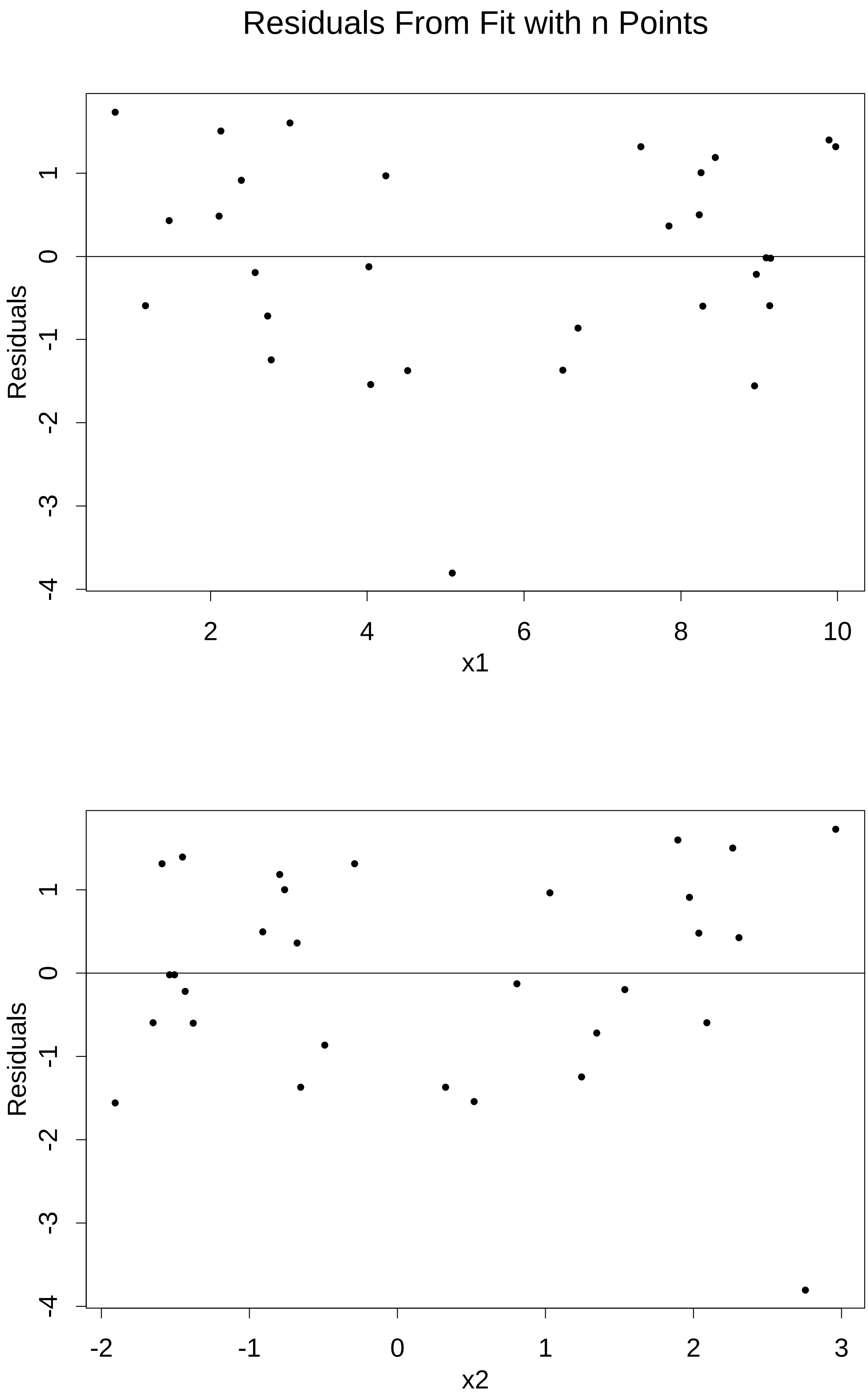
	coef	std.err	t.stat	p.value
Intercept	4.6989	0.3511	13.3843	0
x1	-11.9458	0.0577	-206.9284	0
x2	10.0405	0.1181	85.0499	0

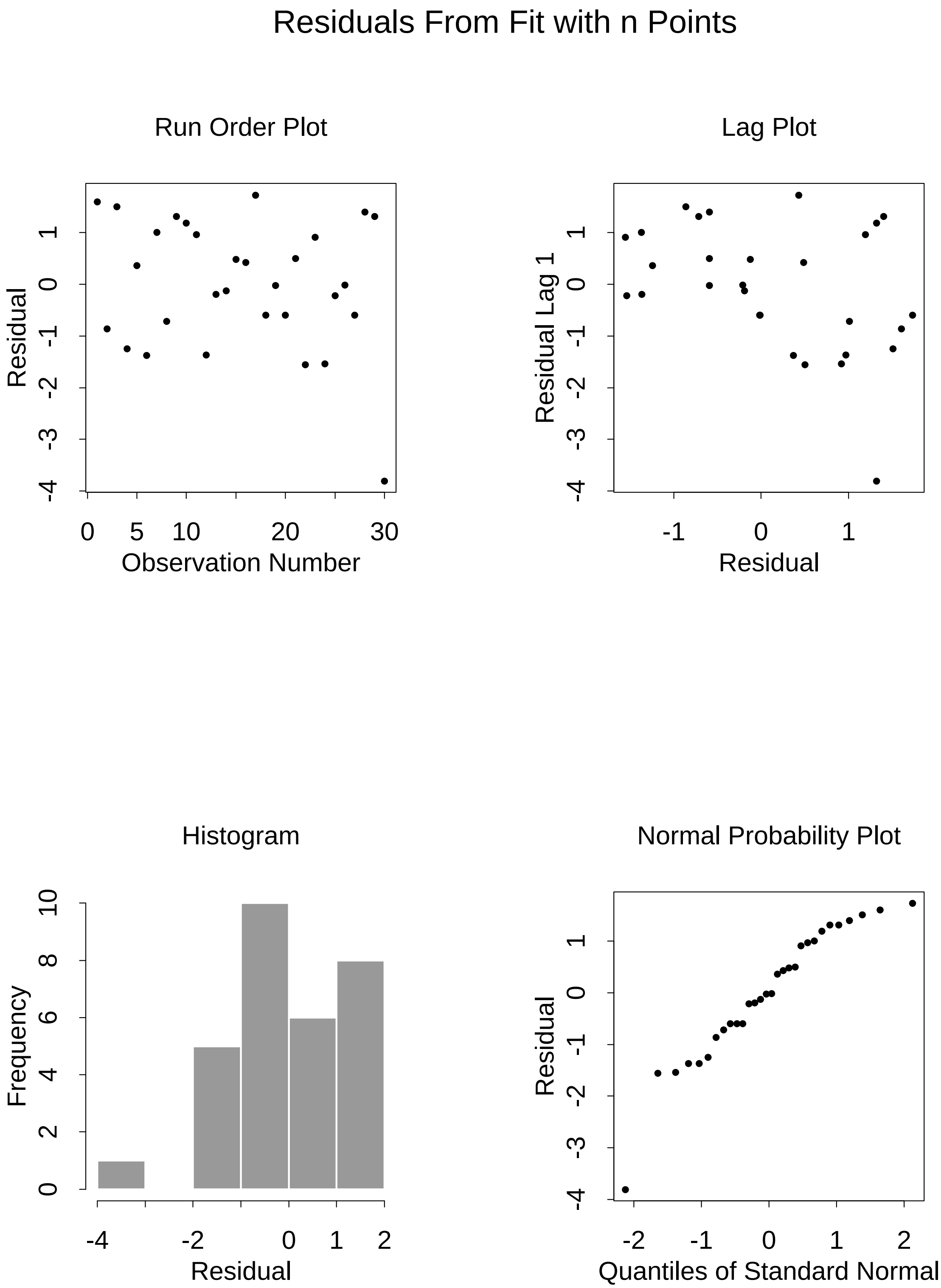

Difference in Predicted Values
from Fits with and without the Leverage Point
Difference in Predicted Values
0.005
0.0
-0.010
-0.020
0
5
10
15
20
25
30
Observation Number

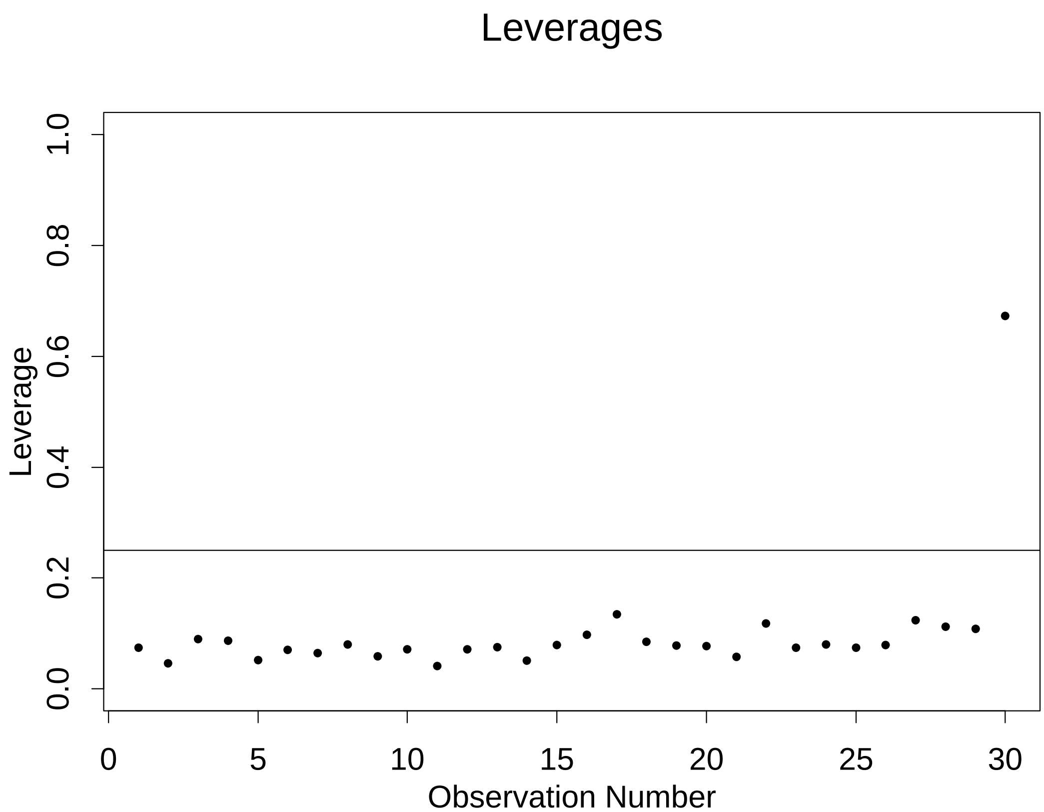

Difference in Predicted Values
on a (Potentially) More Practical Scale
Difference in Predicted Values
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20
25
30
Observation Number

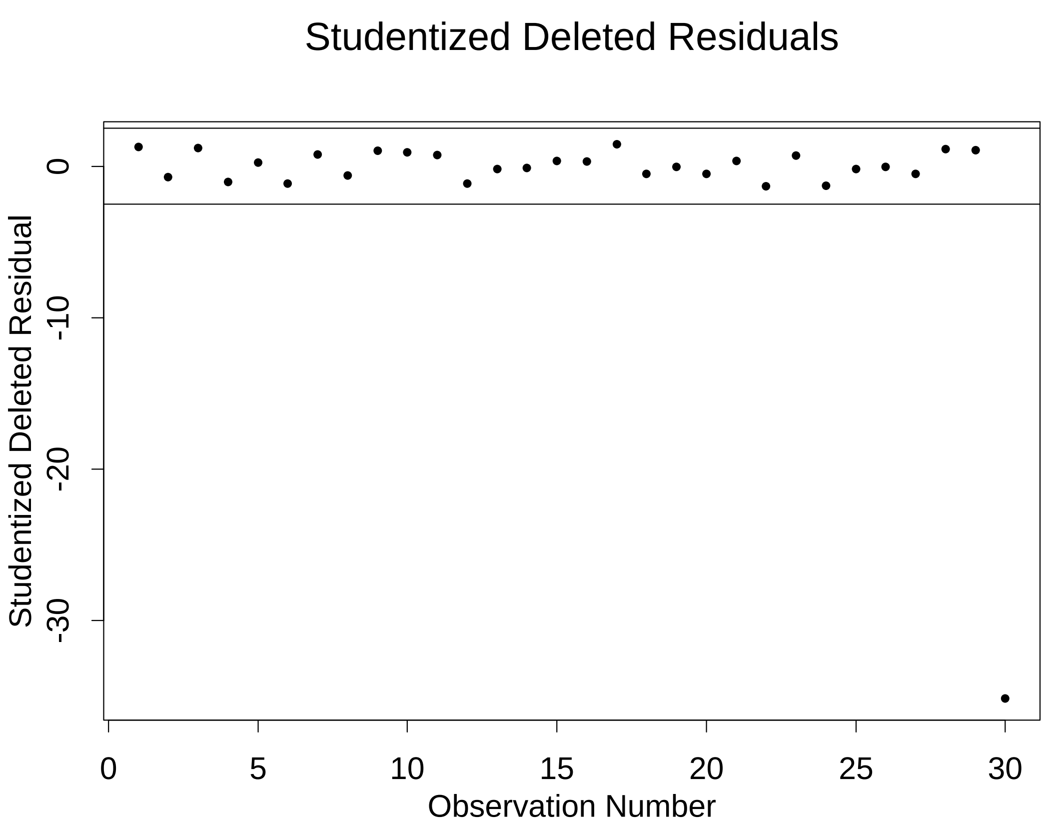
Multivariable Dataset #2 with Potential Outlier

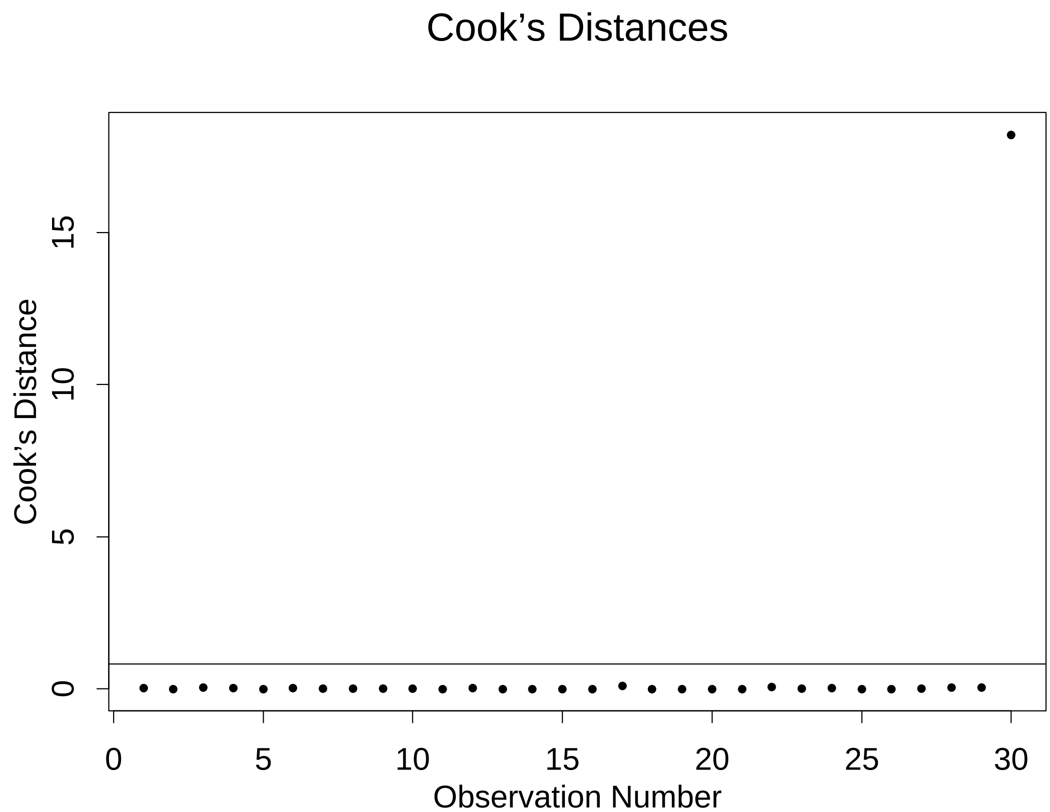


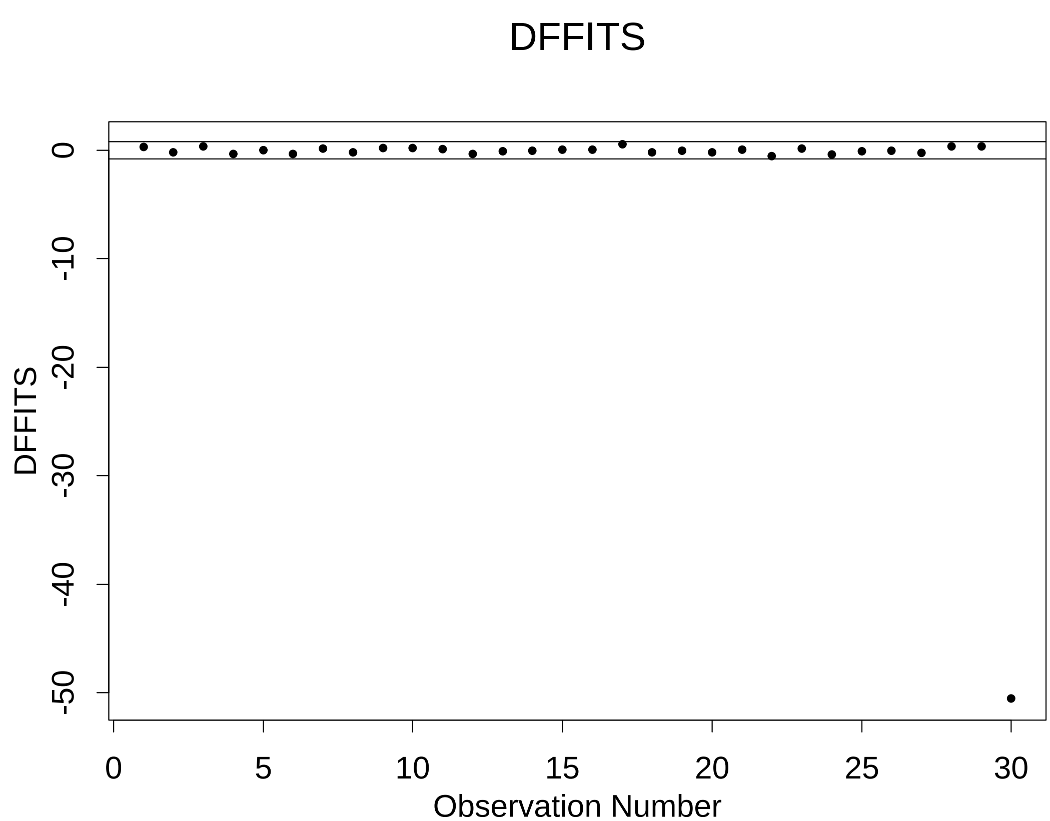

Residuals From Fit with n Points
Residuals
1
0
-1
-2
-3
-4
2
4
6
8
10
x1
Residuals
1
0
-1
-2
-3
-4
-2
-1
0
1
2
3
x2


Residuals From Fit with n Points
Run Order Plot
Residual
Observation Number
Lag Plot
Residual Lag 1
Residual
Histogram
Frequency
Residual
Normal Probability Plot
Residual
Quantiles of Standard Normal


Leverages
Leverage
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Observation Number


Studentized Deleted Residuals
Studentized Deleted Residual
0
-10
-20
-30
0
5
10
15
20
25
30
Observation Number


Cook's Distances
Cook's Distance
0
5
10
15
0
5
10
15
20
25
30
Observation Number

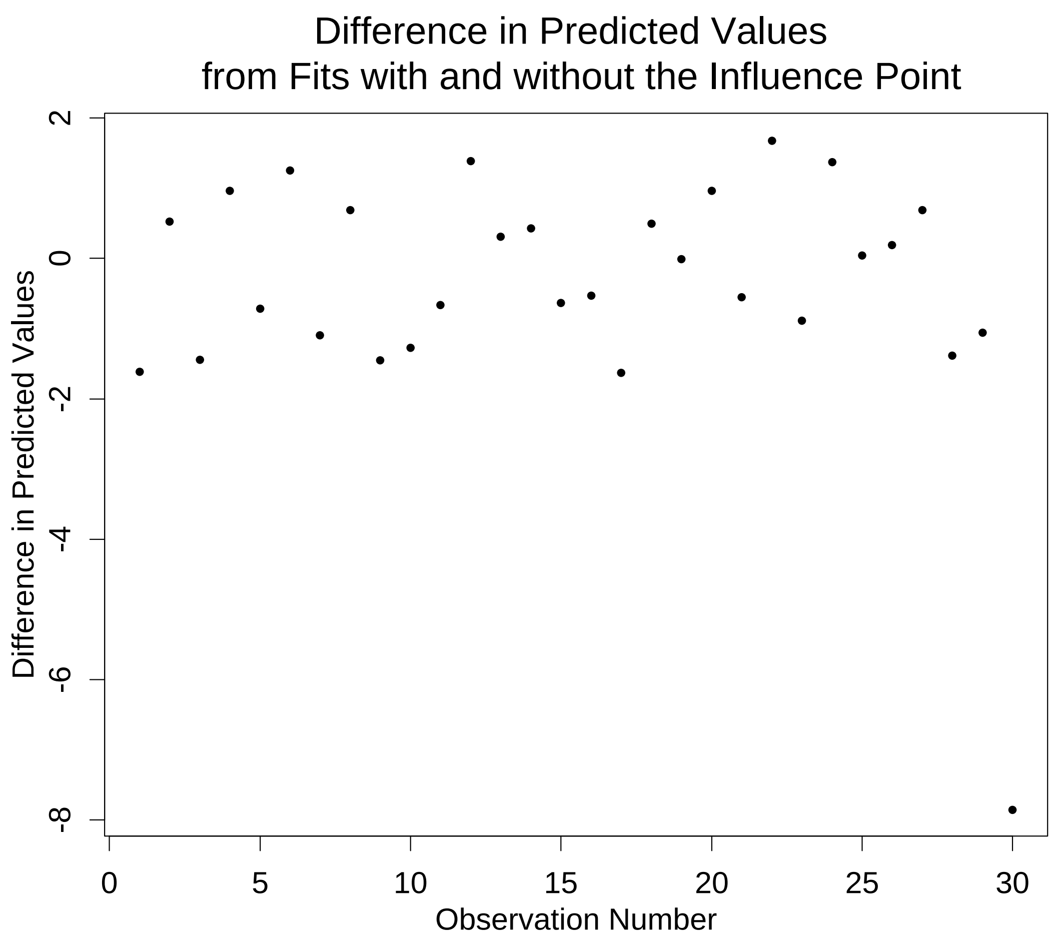

DFFITS
DFFITS
0
-10
-20
-30
-40
-50
0
5
10
15
20
25
30
Observation Number

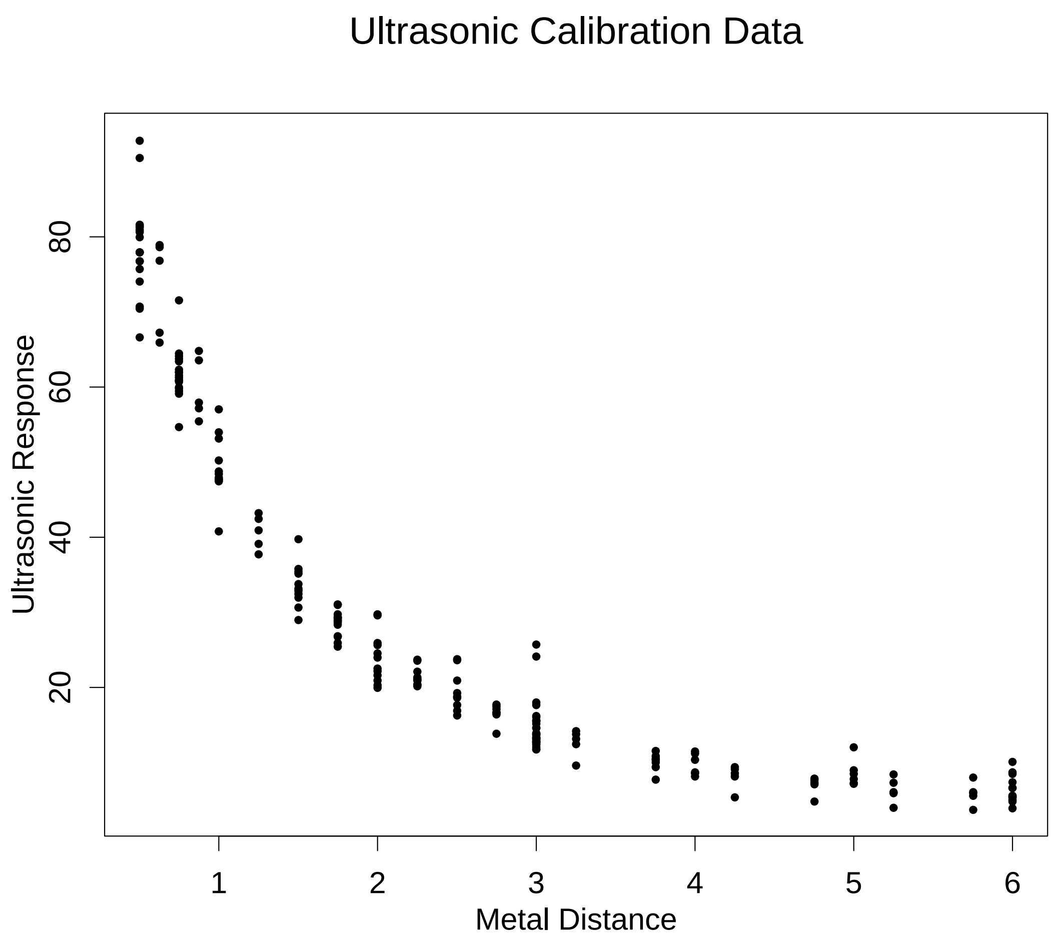
Regression Output Using n Points

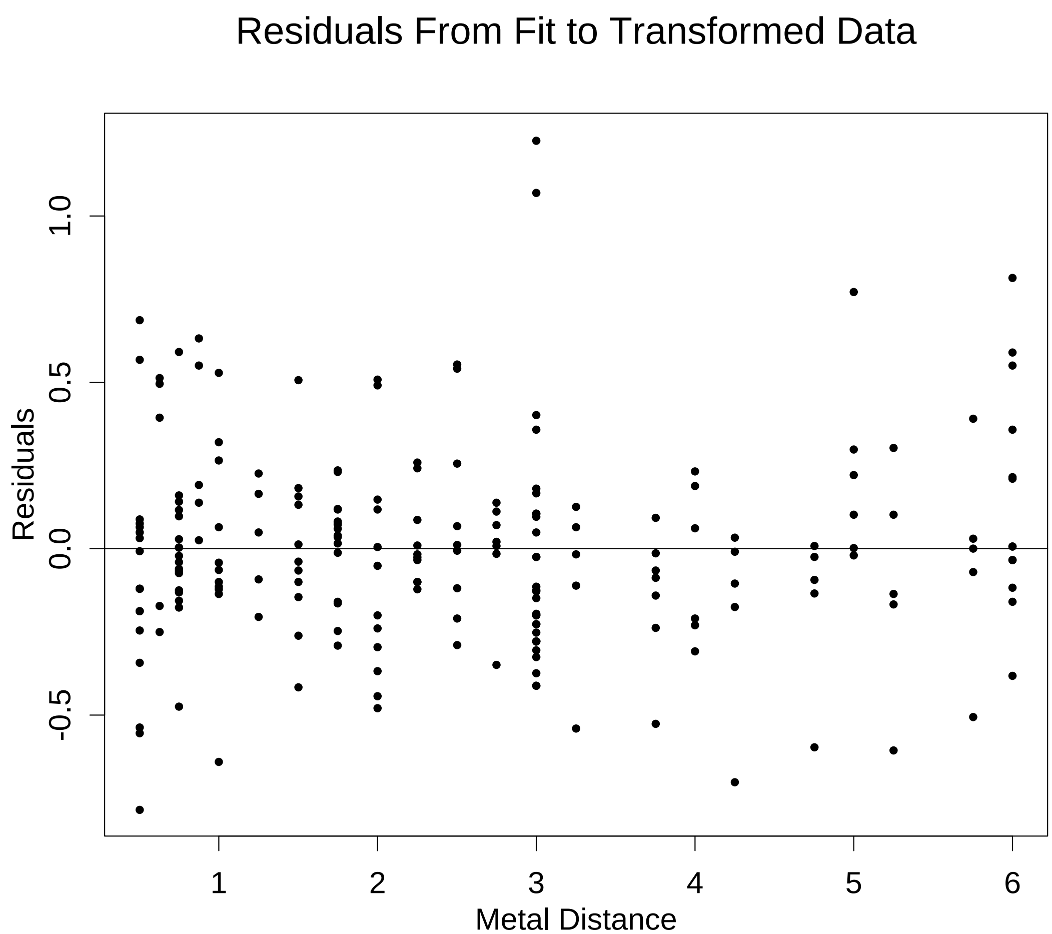
```
N = 30

Residual Standard Error = 1.2939

Multiple R-Square = 0.9994

F-statistic = 22959.24 on 2 and 27 df, p-value = 0

              coef std.err   t.stat p.value
Intercept  14.3252  1.4980   9.5630       0
   x1     -13.5540  0.2401 -56.4605       0
   x2       6.6670  0.4686  14.2261       0
```

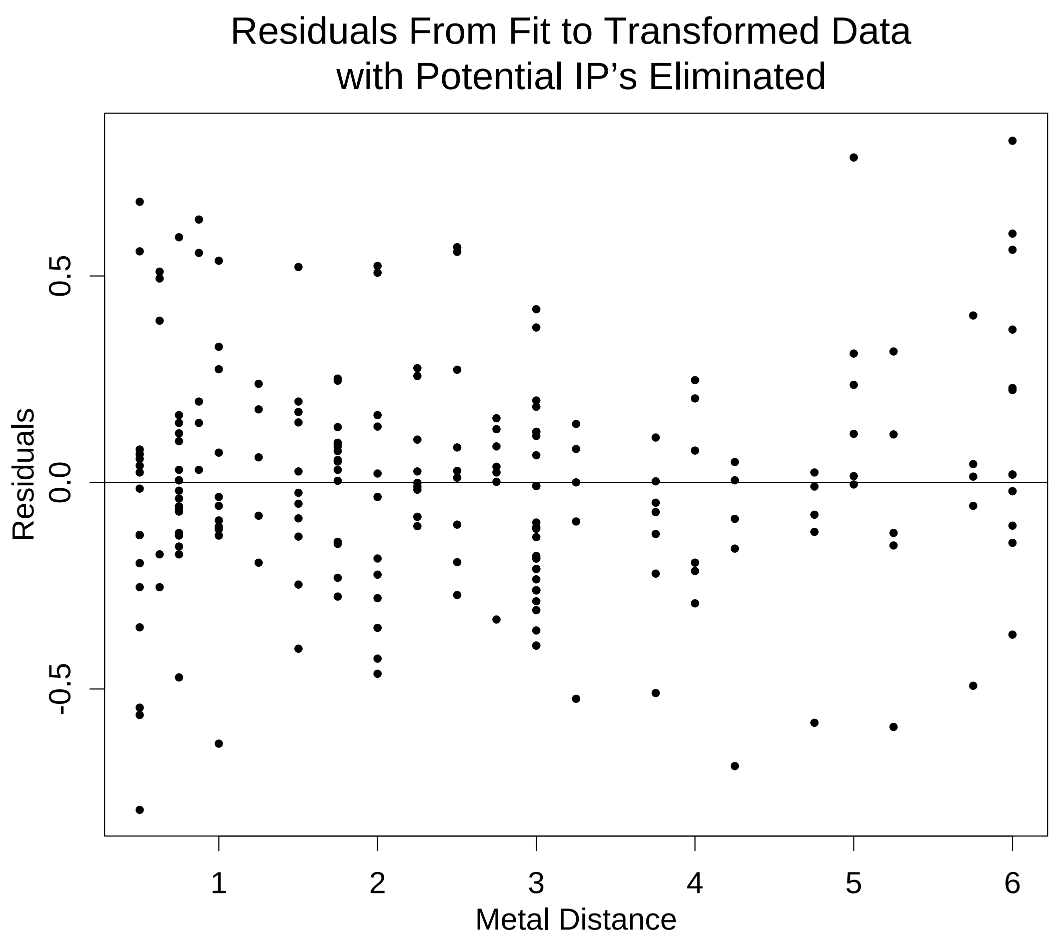
Regression Output Using with the Influential Point Deleted

```
N = 29

Residual Standard Error = 0.1895

Multiple R-Square = 1

F-statistic = 1065271 on 2 and 26 df, p-value = 0

              coef std.err    t.stat p.value
Intercept   4.6989  0.3511   13.3843       0
   x1     -11.9458  0.0577 -206.9284       0
   x2      10.0405  0.1181   85.0499       0
```


Difference in Predicted Values
from Fits with and without the Influence Point
Difference in Predicted Values
2
0
-2
-4
-6
-8
0
5
10
15
20
25
30
Observation Number


Ultrasonic Calibration Data
Ultrasonic Response
20
40
60
80
1
2
3
4
5
6
Metal Distance


Residuals From Fit to Transformed Data
Residuals
1.0
0.5
0.0
-0.5
1
2
3
4
5
6
Metal Distance

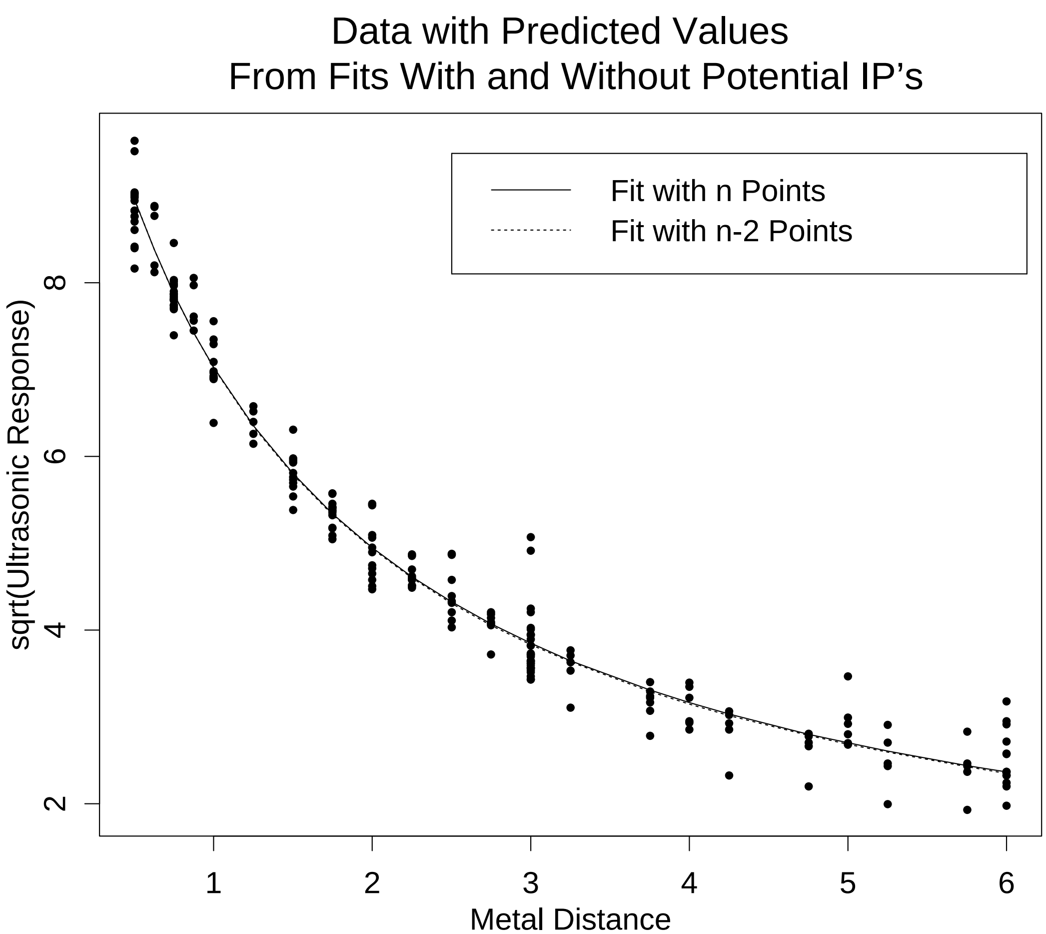

Residuals From Fit to Transformed Data
with Potential IP's Eliminated
Residuals
0.5
0.0
-0.5
1
2
3
4
5
6
Metal Distance

Ultrasonic Data Regression Output: n Points

```
Formula: sqrt(ur) ~ exp( - b1 * md)/(b2 + b3 * md)

Parameters:

        Value Std. Error   t value
b1 -0.0154274 0.00861101  -1.79159
b2  0.0806725 0.00150574  53.57670
b3  0.0638570 0.00288001  22.17250

Residual standard error: 0.29715 on 211 degrees of freedom
```

Ultrasonic Data Regression Output: n-2 Points

```
Formula: sqrt(ur) ~ exp( - b1 * md)/(b2 + b3 * md)

Parameters:

        Value Std. Error   t value
b1 -0.0155610 0.00799912  -1.94533
b2  0.0803004 0.00140502  57.15260
b3  0.0644030 0.00268723  23.96630

Residual standard error: 0.276145 on 209 degrees of freedom
```


Data with Predicted Values
From Fits With and Without Potential IP's
Fit with n Points
Fit with n-2 Points
sqrt(Ultrasonic Response)
2
4
6
8
1
2
3
4
5
6
Metal Distance

Section 2: Summary

Two problems that often occur when carrying out a regression analysis are:

1. finding non-constant standard deviations across different predictor variable values, and

2. finding outliers, high leverage points, and/or influential points in the data.

Outliers, leverage points and influential points can be identified using either graphical residual analysis or more specialized detection methods.

Non-constant standard deviation across the predictors can be corrected using either transformation of the data or weighted least squares.