Statistics for Scientists & Engineers

Regression Models

Session 1 -- Tuesday,   September 12, 2000
Session 2 -- Thursday, September 14, 2000
Session 3 -- Tuesday,   September 19, 2000
Session 4 -- Thursday, September 21, 2000
Session 5 -- Tuesday,   September 26, 2000

Administration Bldg - LR C
9:00 am - 12:00 noon

Will Guthrie
Statistical Engineering Division

will.guthrie@nist.gov
x2854

Course Outline

Section 1: Model Fitting Fundamentals

Section 2: Outliers & Other Problems

Section 3: Prediction & Calibration

Corresponding
Information in Text:

Section 1: § 2.3, 3.1-3.5, 3.7-3.9, 4.1-4.5, 4.7-4.8 and 9.1-9.3

Section 2: § 5.1-5.5, 8.1-8.2 and 9.4

Section 3: § 1.6, 2.2, 3.6, 4.6, and 6.1-6.4

Section Outline

1. Definitions, Concepts & Assumptions
2. Selection of the Regression Function
3. Estimation of Model Parameters
4. Model Validation
5. Additional Examples

Regression

Regression analysis is the concise description of multivariate data by partitioning it into a deterministic component given by a mathematical function and a random component which follows a probability distribution.

Data

The multivariate data used for regression consists of:

1. a ‘response variable’, y, which is also called a ‘dependent variable’, and

2. one or more ‘predictor variables’,

   $$x_1, x_2, \ldots, x_k,$$

   which are also called ‘independent variables’.

Model

The description of the data resulting from a regression analysis is called ‘the model’ of the data.

In general, the model is written:

$$y = f(x_1, x_2, \ldots, x_k; \beta_1, \beta_2, \ldots, \beta_p) + \varepsilon.$$

Model

Some examples of specific regression models include:

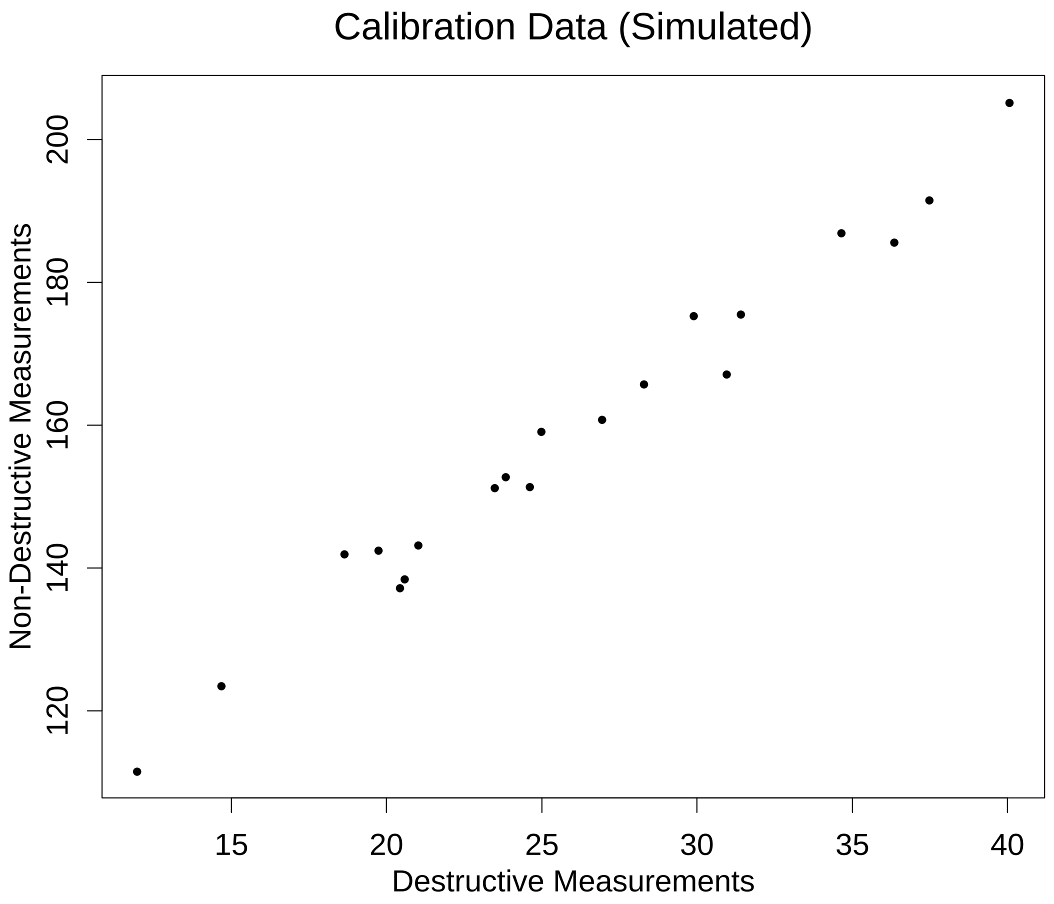
$$y = \beta_1 + \beta_2 x + \varepsilon$$

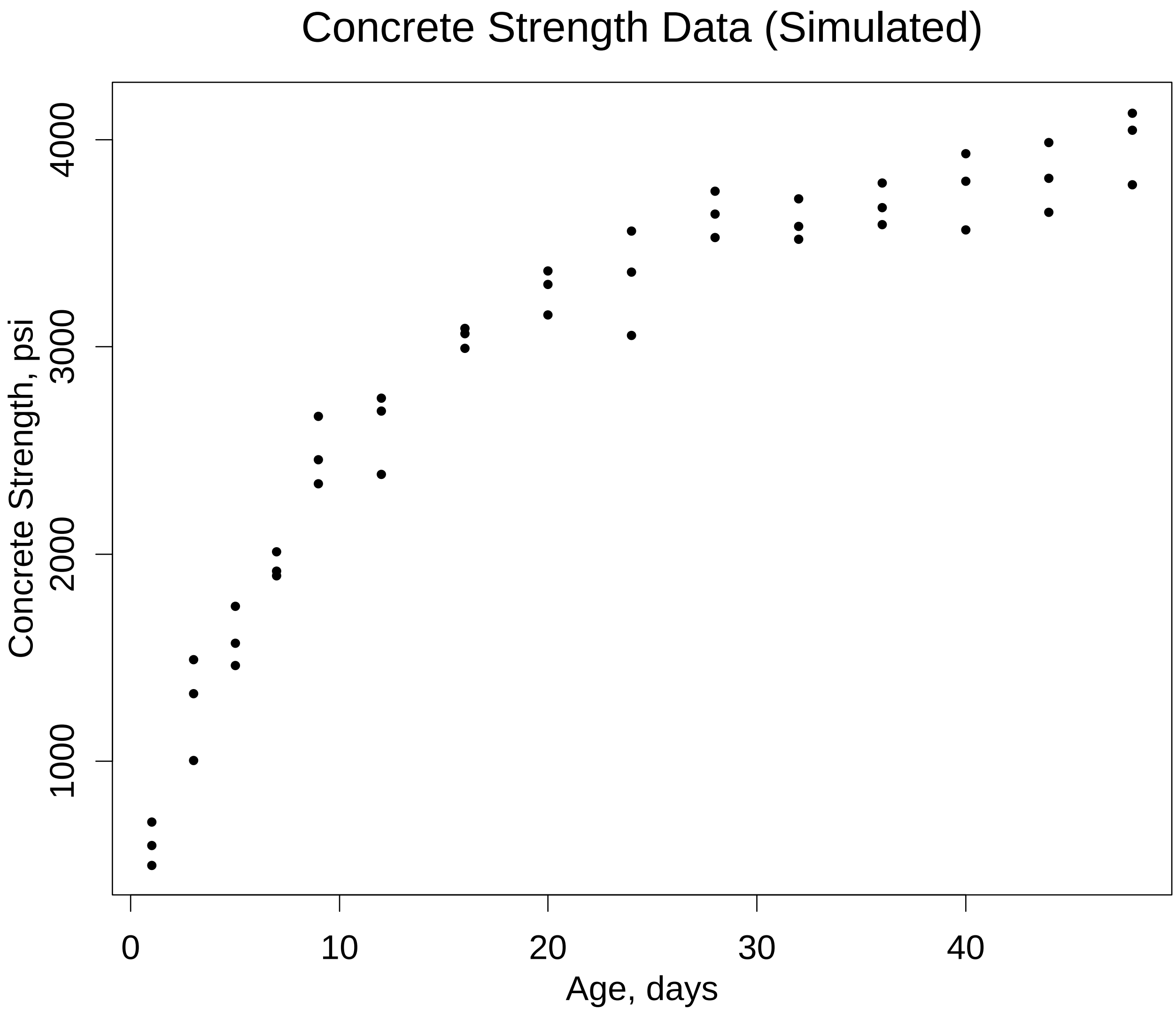
$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \varepsilon$$

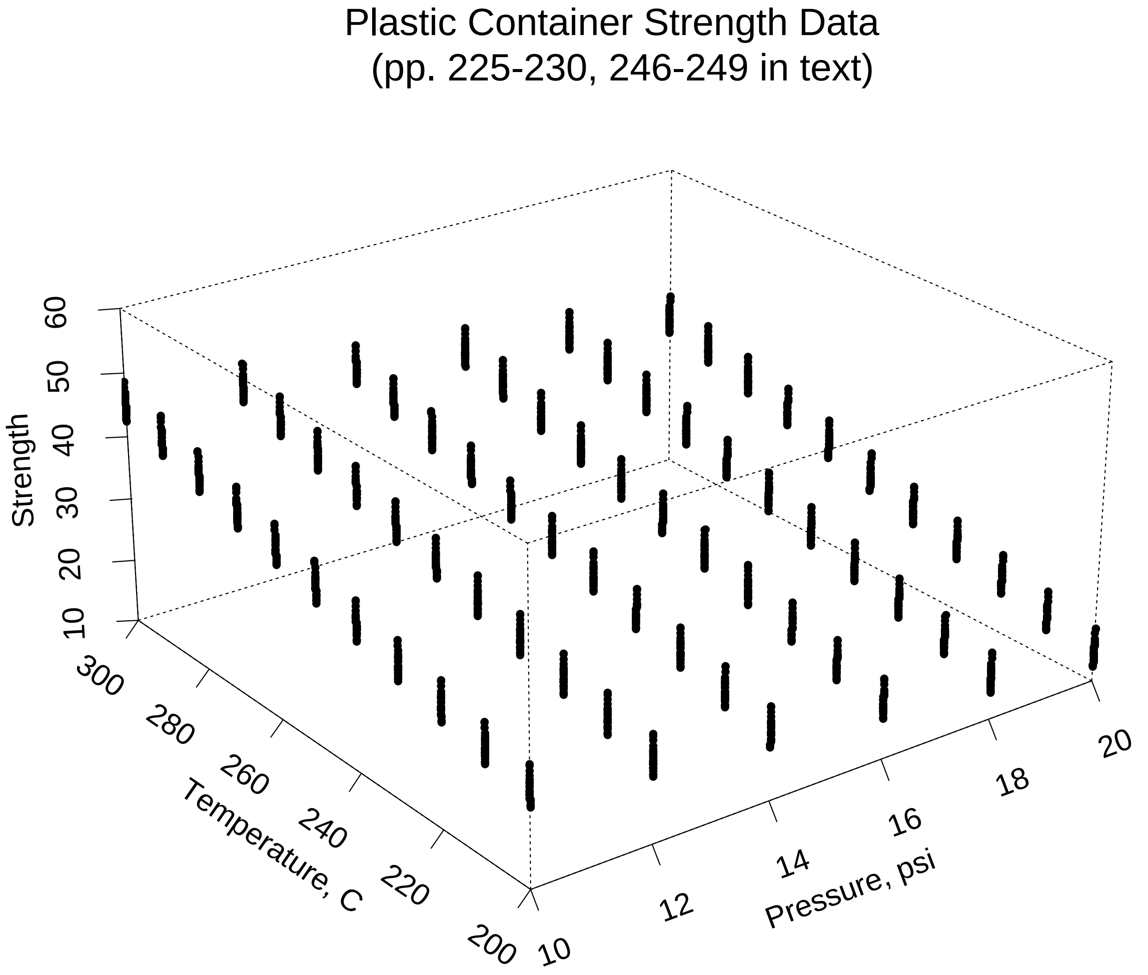
$$y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1 x_2 + \varepsilon$$

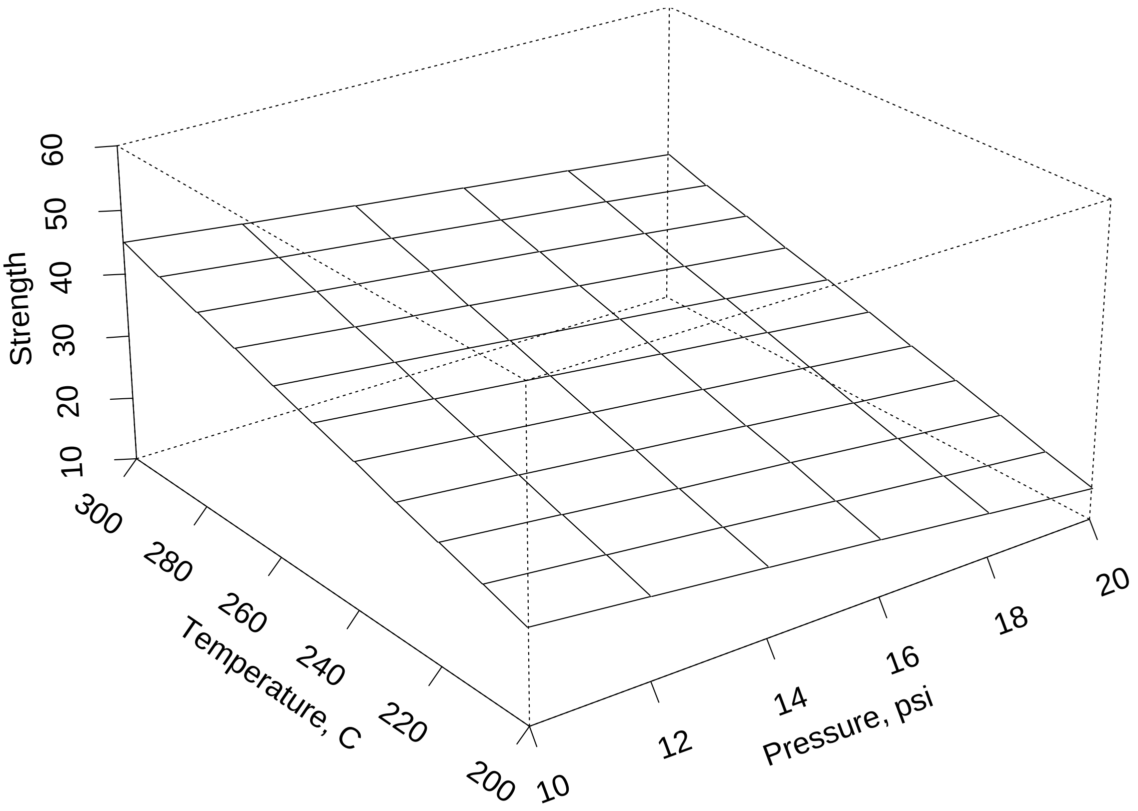
$$y = \frac{\beta_1 + \beta_2 x}{1 + \beta_3 x} + \varepsilon$$

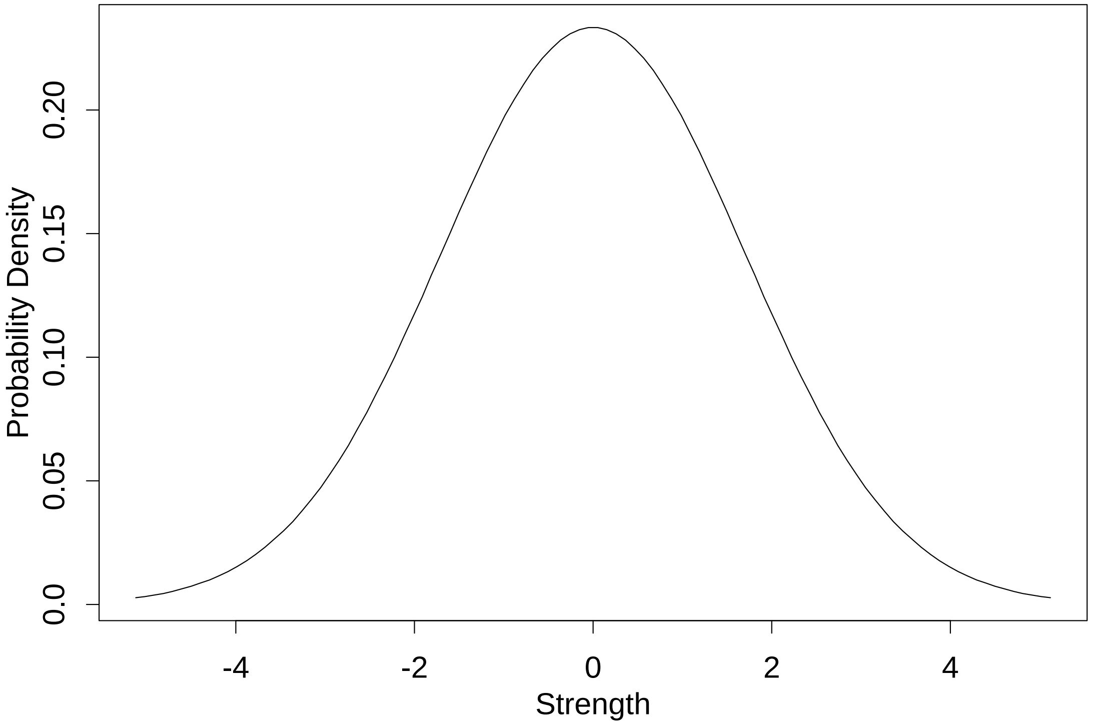
$$y = \beta_1 \exp(-\beta_2 x) + \varepsilon$$


Calibration Data (Simulated)
Non-Destructive Measurements
120
140
160
180
200
15
20
25
30
35
40
Destructive Measurements


Concrete Strength Data (Simulated)
Concrete Strength, psi
1000
2000
3000
4000
0
10
20
30
40
Age, days


Plastic Container Strength Data
(pp. 225-230, 246-249 in text)
Strength
10
20
30
40
50
60
300
280
260
240
220
200
Temperature, C
10
12
14
16
18
20
Pressure, psi


Strength
10
20
30
40
50
60
Temperature, C
300
280
260
240
220
200
Pressure, psi
10
12
14
16
18
20


Probability Density
0.0
0.05
0.10
0.15
0.20
-4
-2
0
2
4
Strength

Assumptions about the Data

1. The data actually follow a population model of the type $y = f(x_1, \dots, x_k, \beta_1, \dots, \beta_p) + \varepsilon$.

2. The complete observations, $(x_1, x_2, \dots, x_k, y)$, are randomly sampled from the population model, or the y's are randomly sampled for a set of preselected values of $(x_1, x_2, \dots, x_k)$.

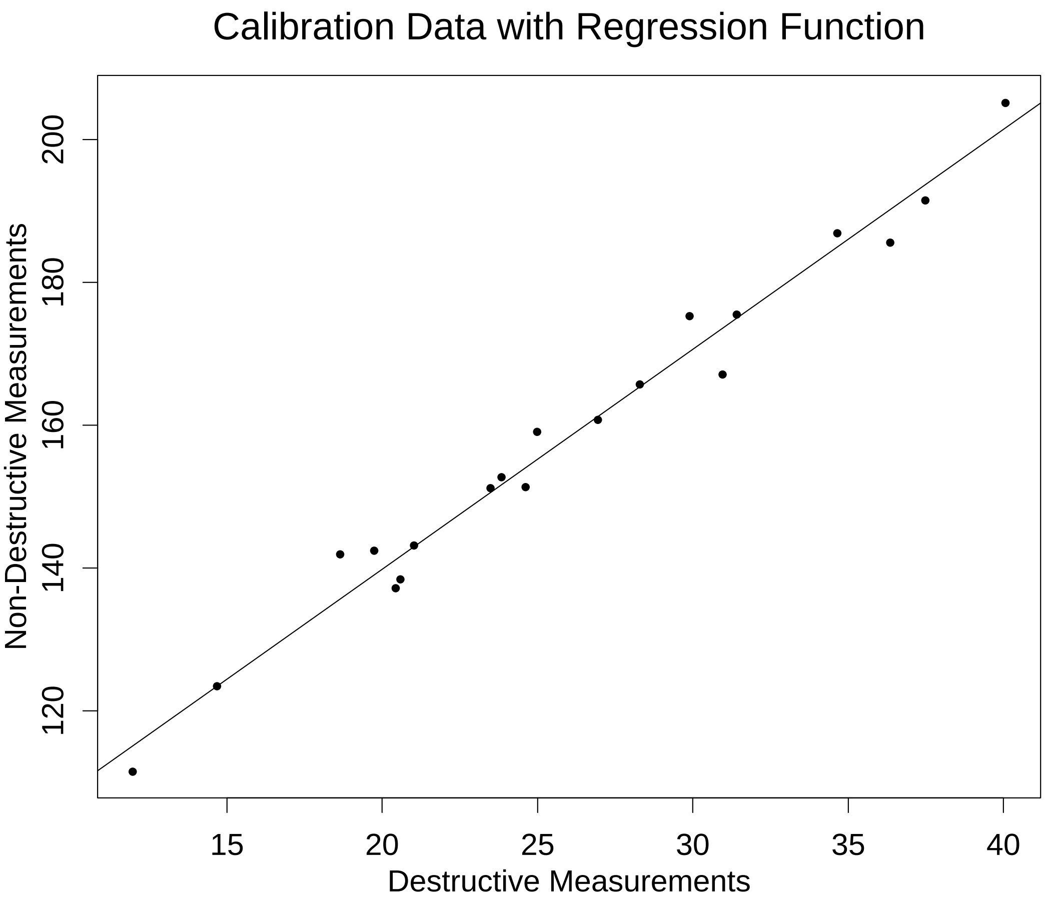
3. The predictor variables, $x_1, x_2, \dots, x_k$, are measured or observed without error.

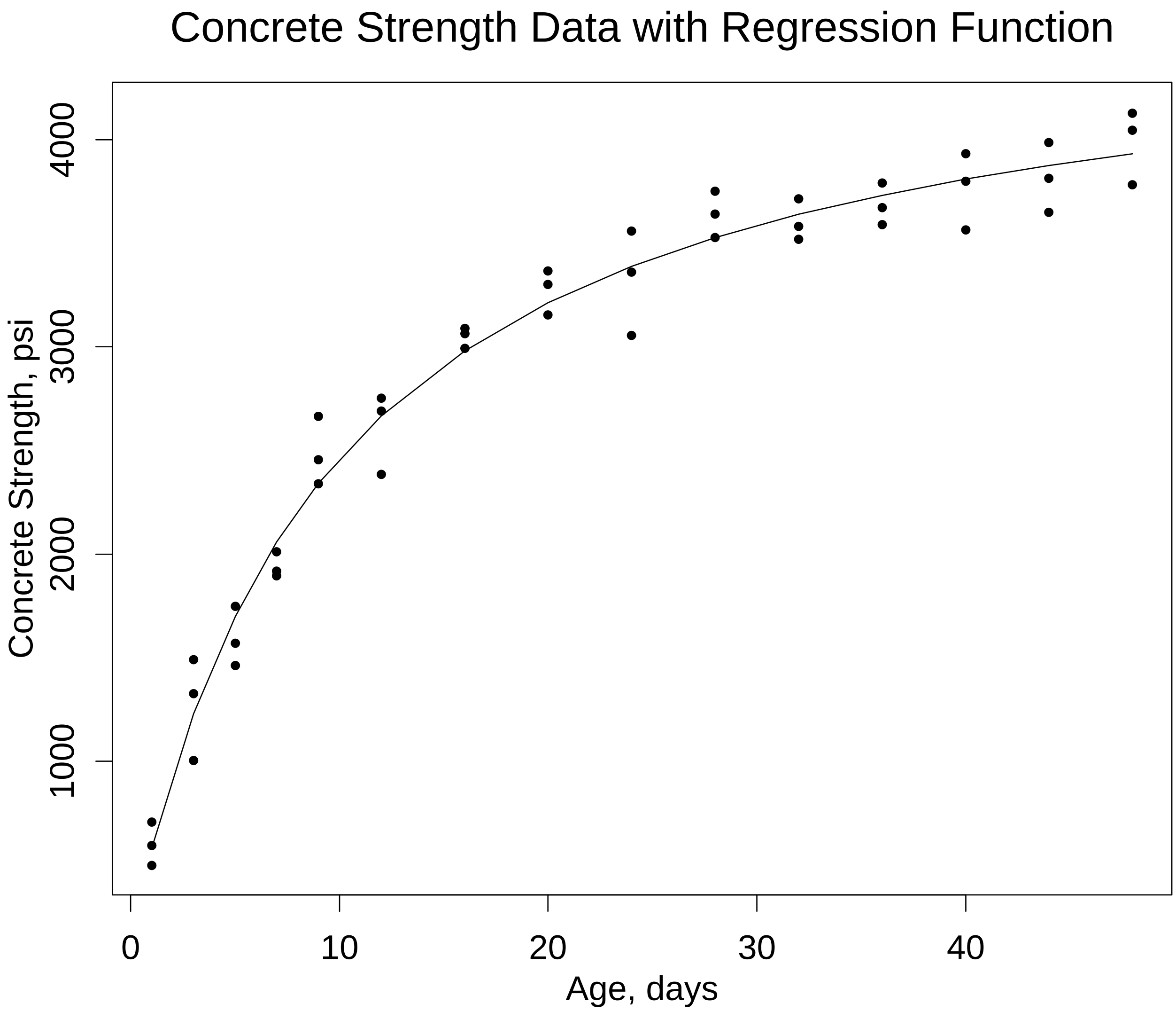
Assumptions about the Model

1. The mean, μ, of the random errors is zero for each combination of predictor variable values.

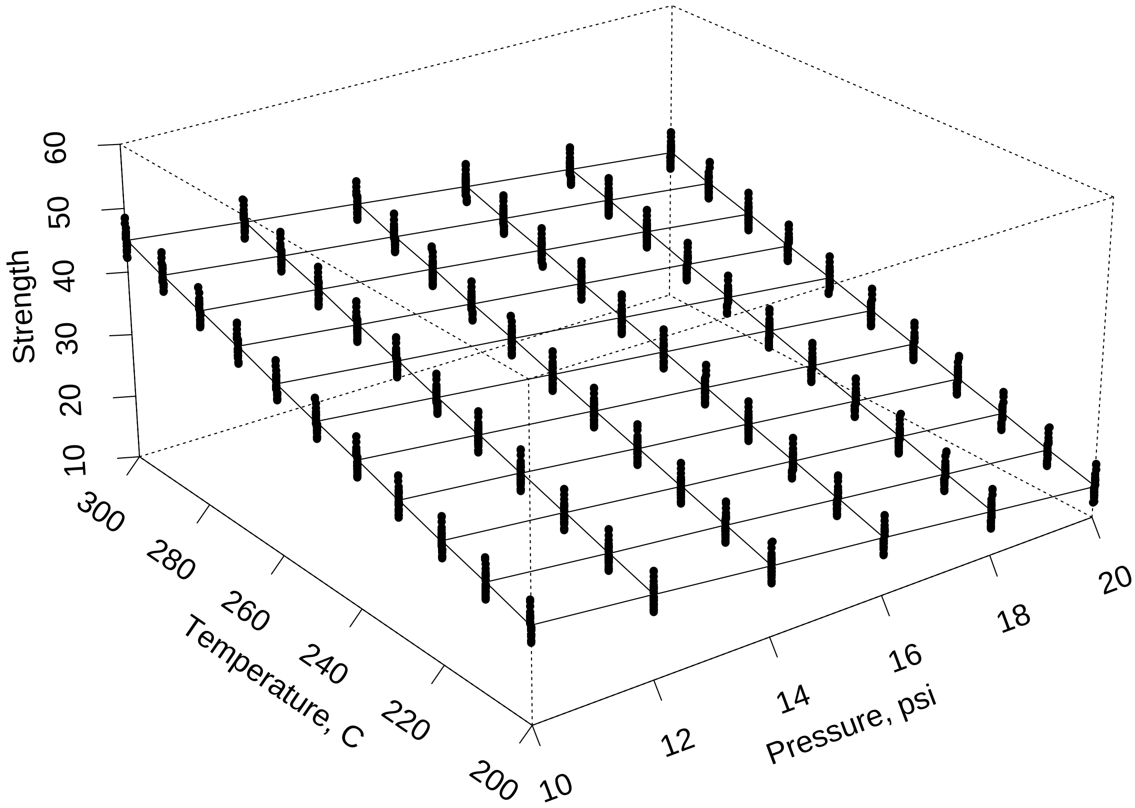
2. The standard deviation, σ, of the random errors is constant for each combination of predictor variable values.

3. The random errors follow a normal distribution for each combination of predictor variable values.


Calibration Data with Regression Function
Non-Destructive Measurements
120
140
160
180
200
15
20
25
30
35
40
Destructive Measurements


Concrete Strength Data with Regression Function
Concrete Strength, psi
1000
2000
3000
4000
0
10
20
30
40
Age, days

Plastic Container Strength Data with Regression Function

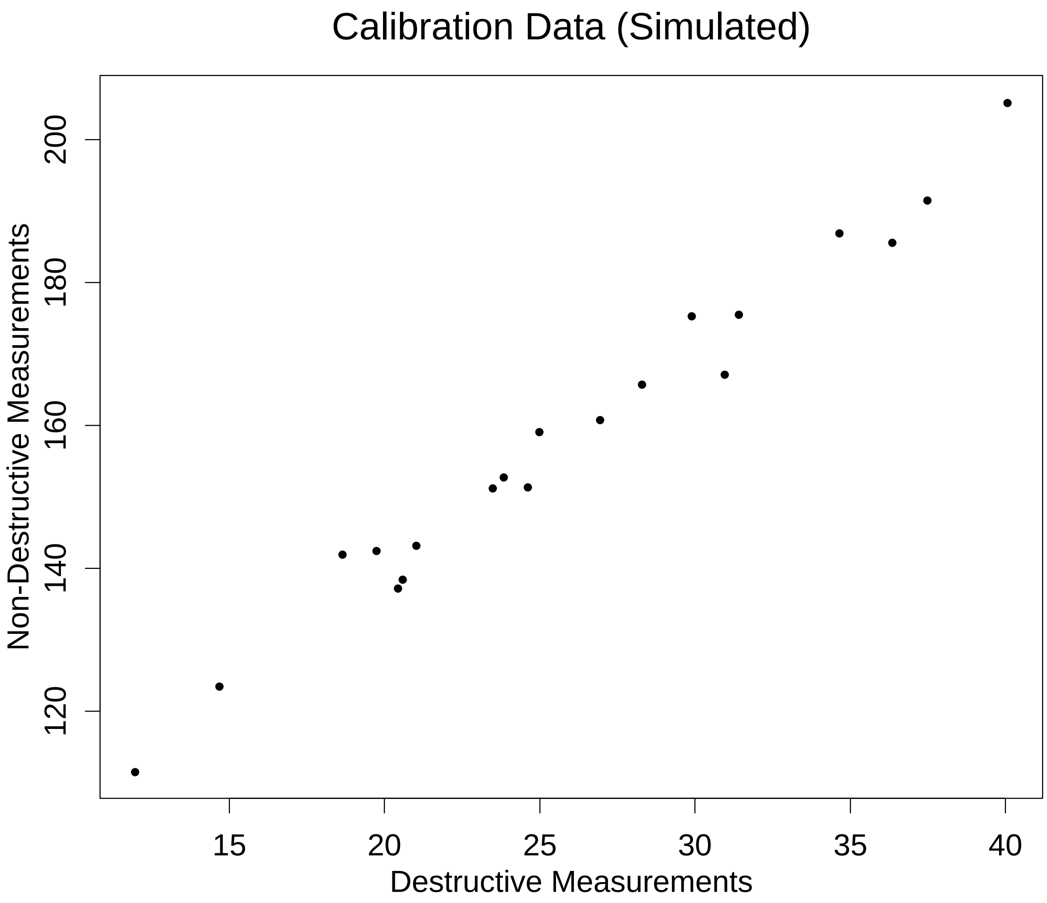


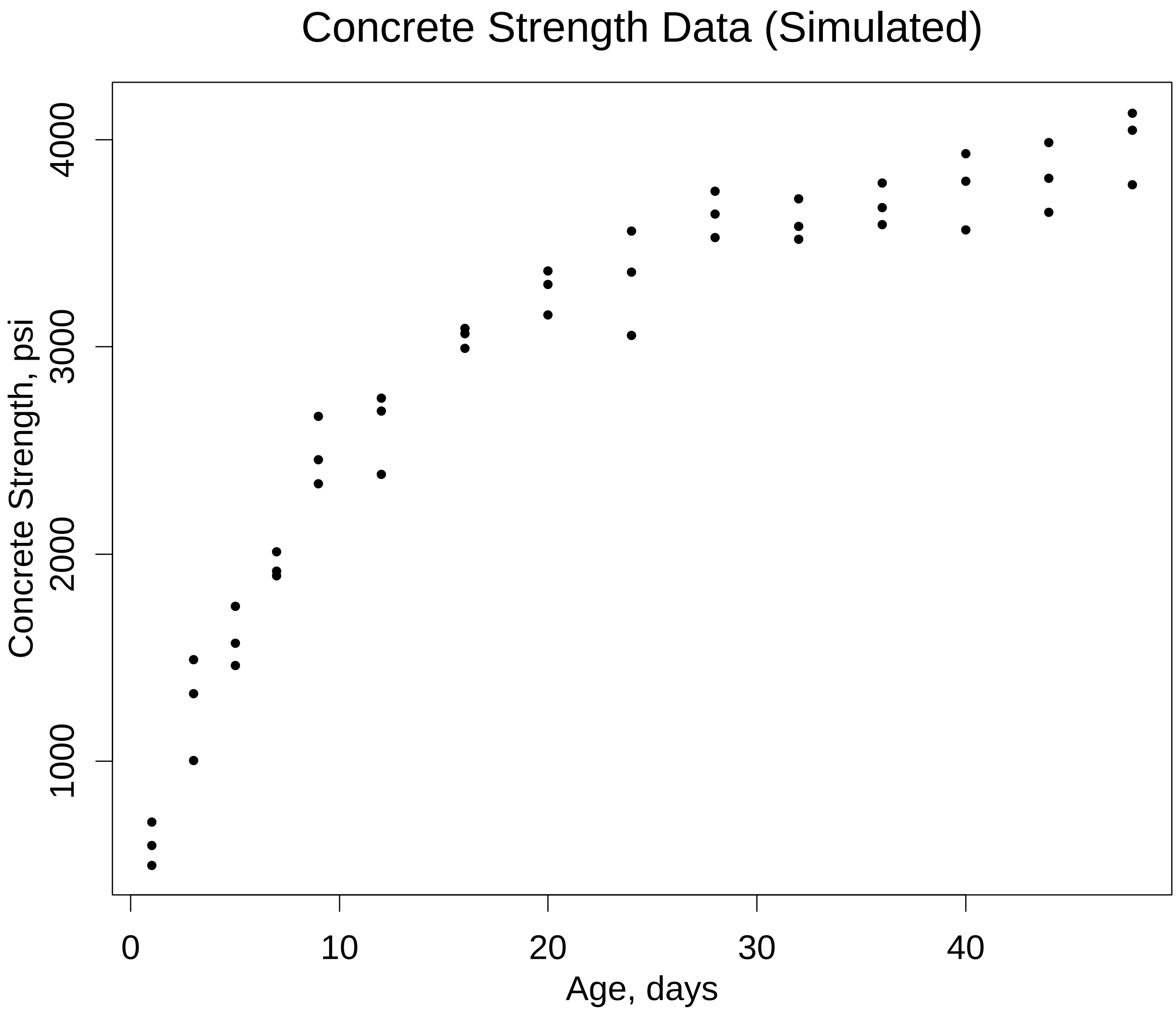
Selection of Regression Function

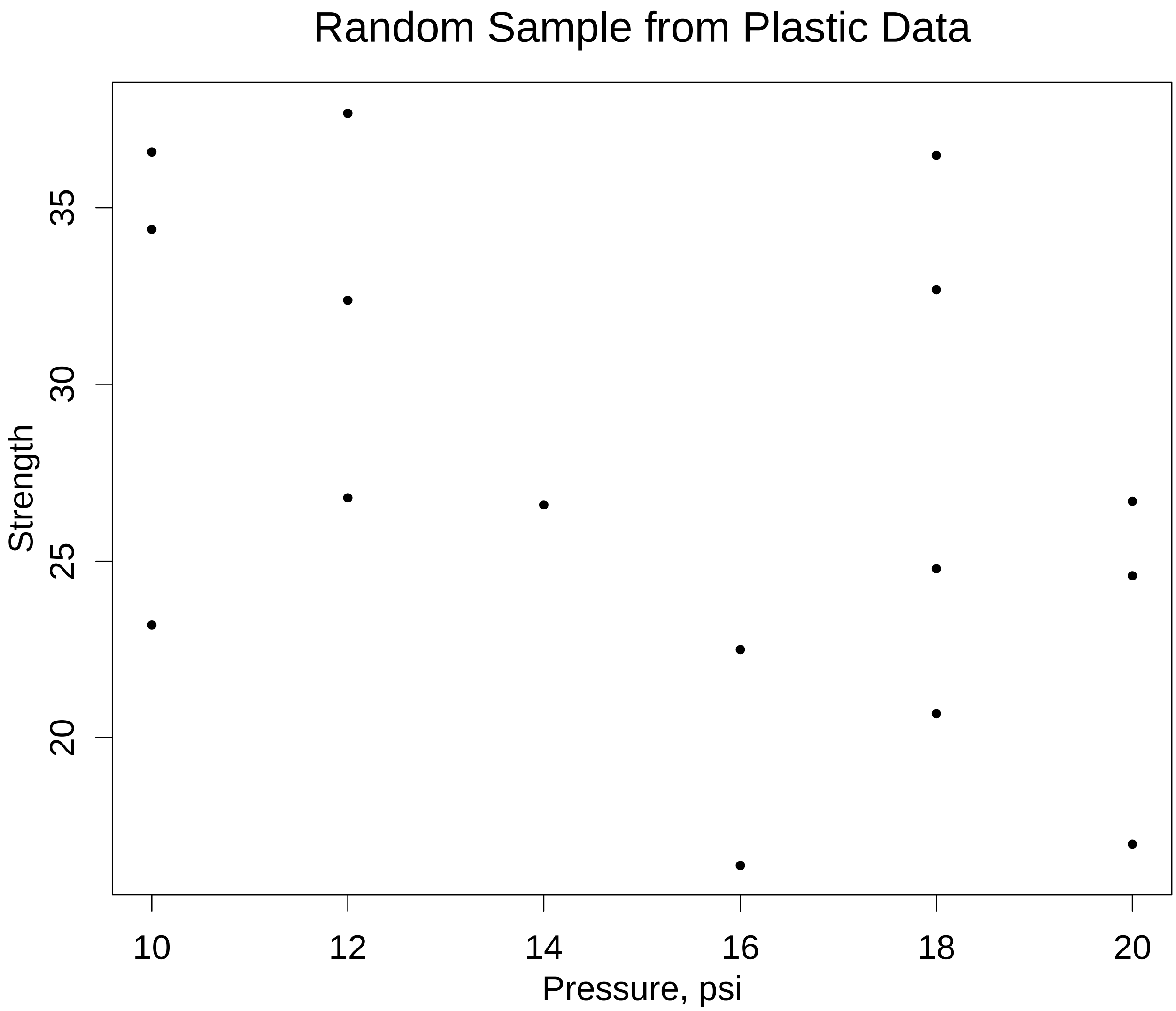
The basic steps for determining the form of the regression function are:

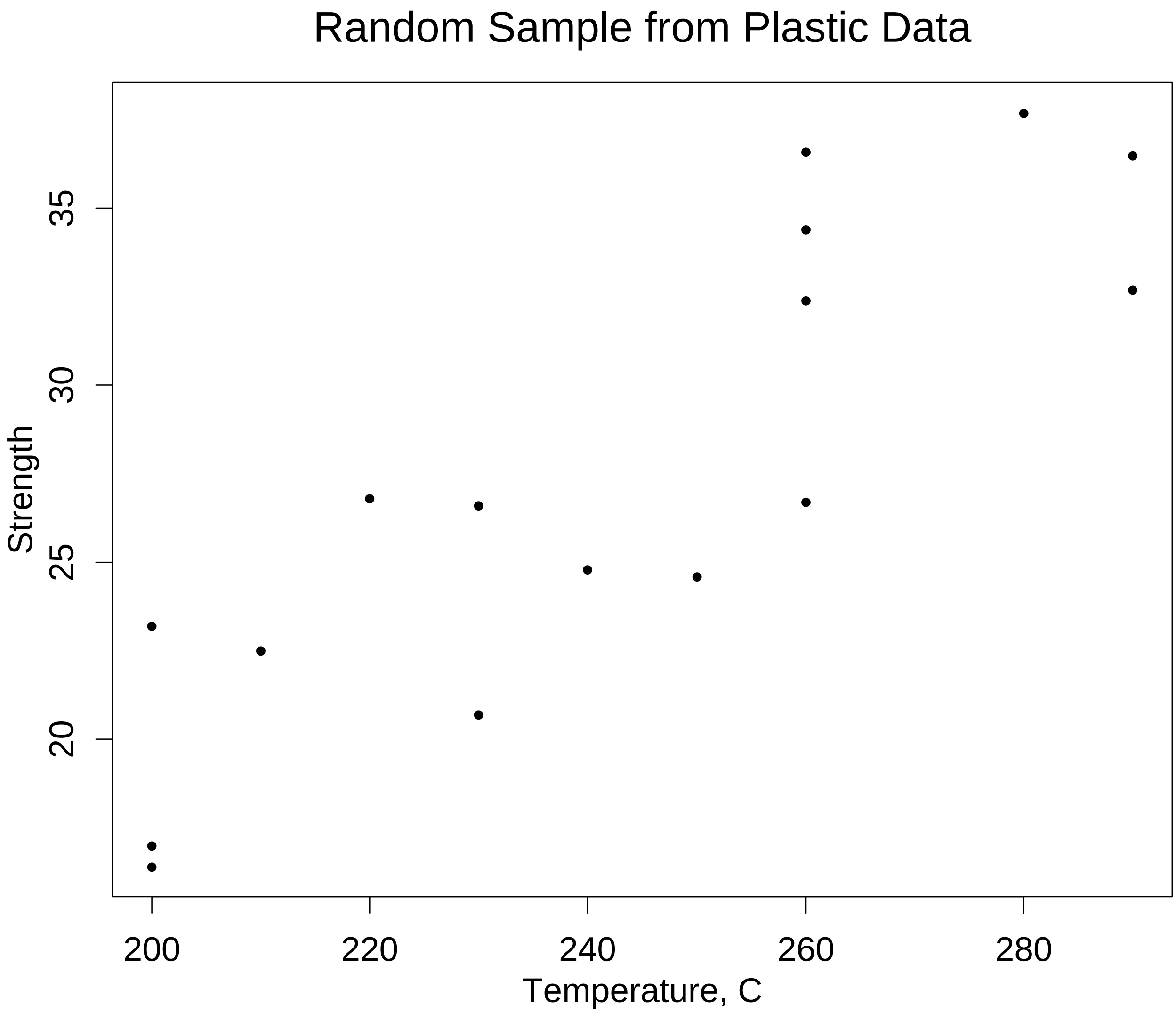
1. Plot the data to confirm the appropriateness of a theoretical function or to determine what rough shape an empirical function should have.

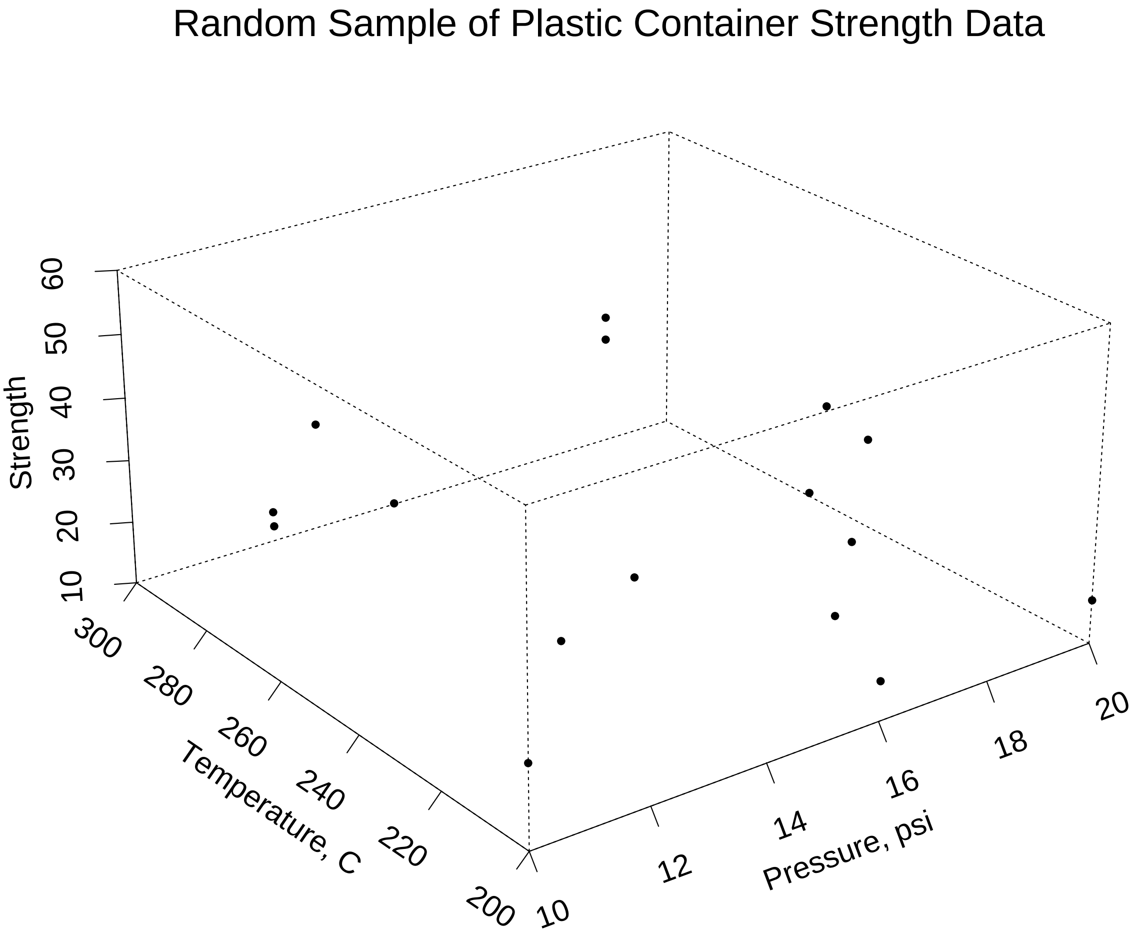
2. Use other scientific knowledge, relevant to the data, to refine the form of the function.

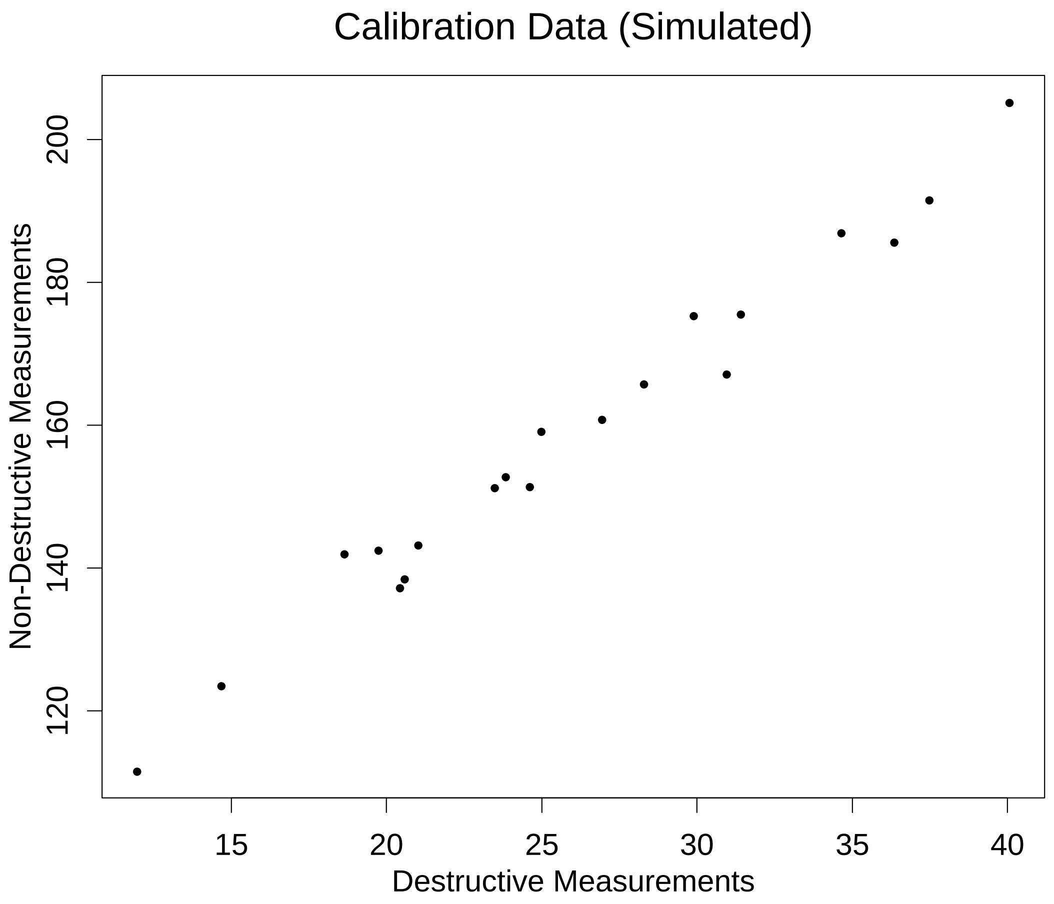

Calibration Data (Simulated)
Non-Destructive Measurements
120
140
160
180
200
15
20
25
30
35
40
Destructive Measurements


Concrete Strength Data (Simulated)
Concrete Strength, psi
1000
2000
3000
4000
0
10
20
30
40
Age, days


Random Sample from Plastic Data
Strength
20
25
30
35
10
12
14
16
18
20
Pressure, psi


Random Sample from Plastic Data
Strength
20
25
30
35
200
220
240
260
280
Temperature, C


Random Sample of Plastic Container Strength Data


Calibration Data (Simulated)
Non-Destructive Measurements
120
140
160
180
200
15
20
25
30
35
40
Destructive Measurements

Least Sum of Squares Estimation

The ‘least sum of squares’ method of estimation provides a way to find good, objective estimates of the true, unknown parameters, $\beta_1, \beta_2, \ldots, \beta_p$.

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

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

Note: $n =$ the number of observations in the data set. The index i in Q refers to the i^{th} individual observation.

LSS Estimates for the Straight Line Regression Function

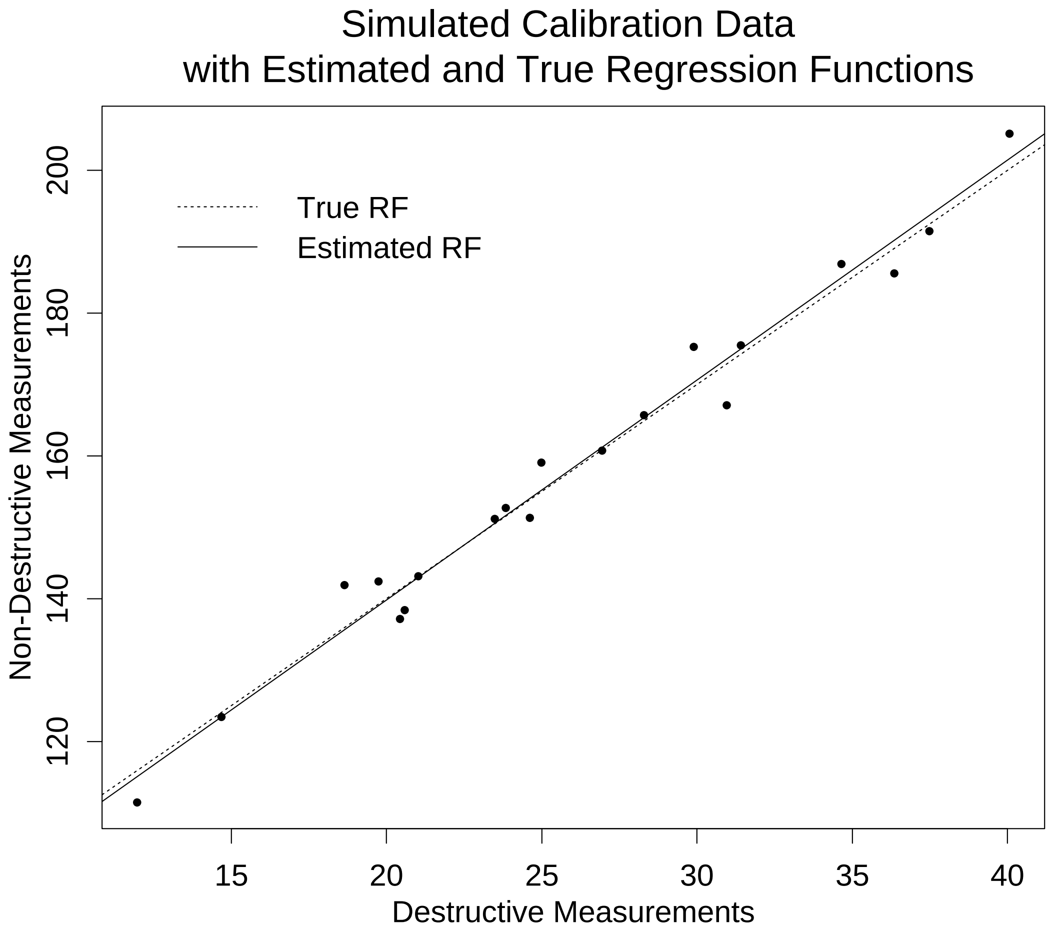
For straight line regression:

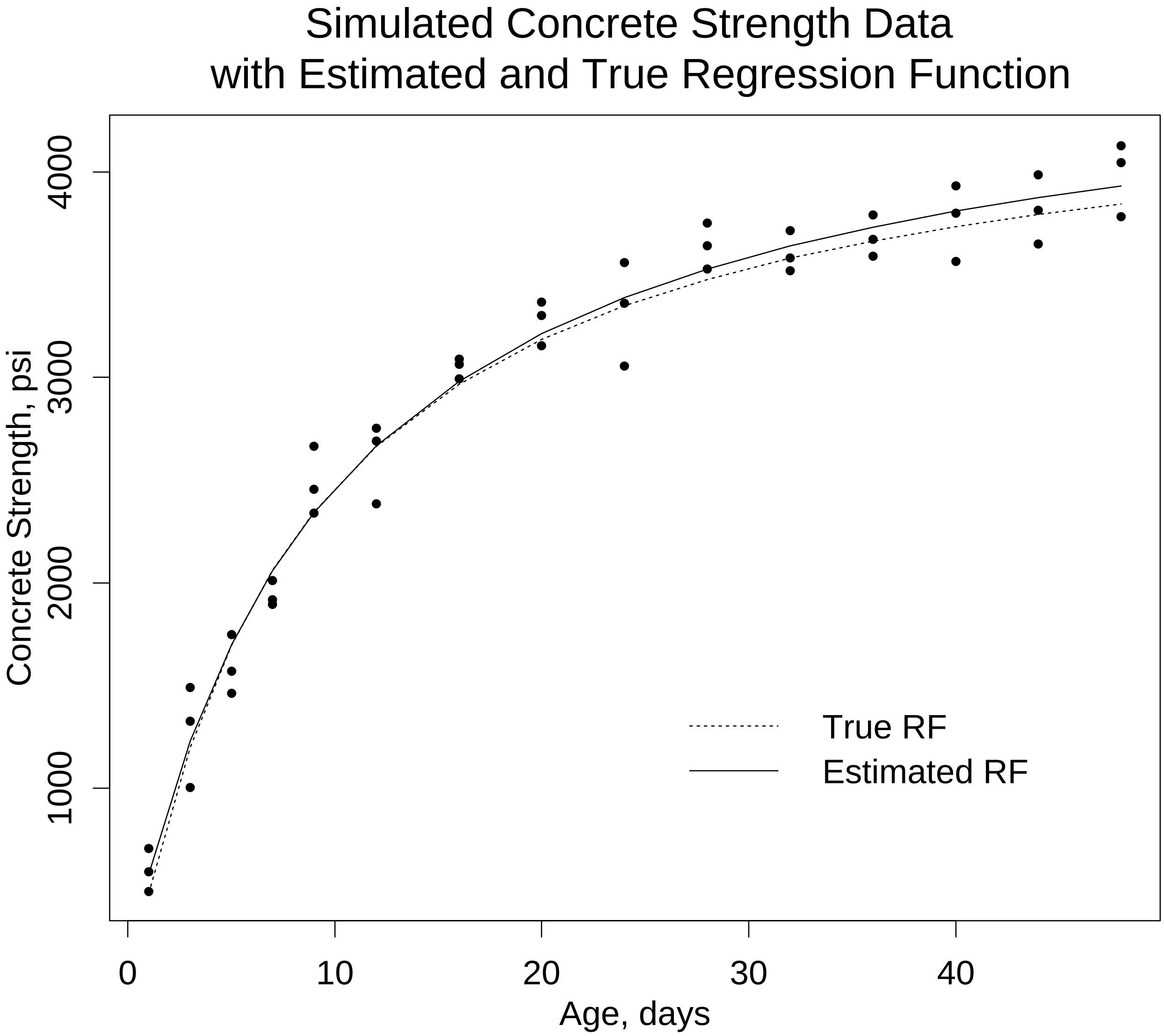
$$Q = \sum_{i=1}^{n} [y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i)]^2$$

After going through the calculations, the formulas for the least squares estimates of the parameters simplify to

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

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


Simulated Calibration Data
with Estimated and True Regression Functions
True RF
Estimated RF
Non-Destructive Measurements
120
140
160
180
200
15
20
25
30
35
40
Destructive Measurements


Simulated Concrete Strength Data
with Estimated and True Regression Function
Concrete Strength, psi
1000
2000
3000
4000
0
10
20
30
40
Age, days
True RF
Estimated RF

Parameter Estimation for Nonlinear Functions

Regression problems can be divided into two basic classes, ‘linear’ and ‘nonlinear’ regression.

Different software is required compute least squares parameter estimates for linear and nonlinear functions. An analytical solution to the LSS minimization exists for linear functions. but, parameter estimates must generally be computed iteratively for nonlinear models.

In addition, nonlinear regression software requires the user to provide starting values for the parameter estimates.

Classification of Regression Problems

The ‘linear’ in linear regression refers to the fact that the regression function is a linear combination of the unknown parameters.

A linear combination of the parameters is a function of the parameters that involves only multiplying each parameter by a constant and/or adding a constant to each.

These regression functions are all ‘linear’, even though they are not straight lines:

$$y = \beta_1 + \beta_2 x + \varepsilon$$

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \varepsilon$$

$$y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1 x_2 + \varepsilon$$

‘Nonlinear’ Regression Functions

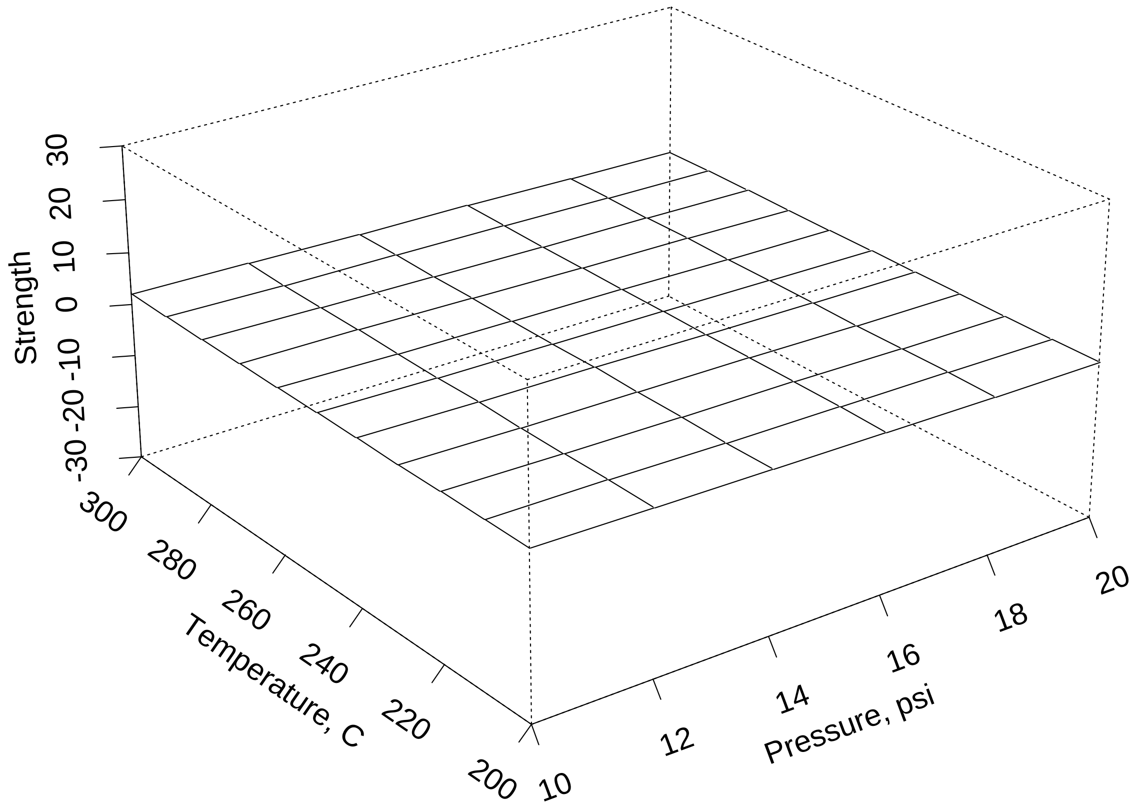
These regression functions are ‘nonlinear’ because they are nonlinear functions of the parameters:

$$y = \frac{\beta_1 + \beta_2 x}{1 + \beta_3 x} + \varepsilon$$

$$y = \beta_1 \exp(-\beta_2 x) + \varepsilon$$

$$y = \beta_1 + \beta_2 \sin(\beta_4 + \beta_5 x) + \varepsilon$$

Difference of Estimated and True Regression Functions
for Random Sample of Plastic Data



Estimation of σ
Motivation Using a Single Population

To estimate the variability of a random variable from a single population, the sample standard deviation is used.

$$s = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

The divisor $n-1$ is used because after estimating the mean, μ from the data, there are only $n-1$ unconstrained deviations from the sample mean left to estimate σ.

Estimation of σ
Extension to Regression

A standard deviation is also used to summarize the random variability in regression data.

However, regression data involves a family of distributions, indexed by the x's, so the sample mean, μ, is replaced by the means associated with the different combinations of $(x_1, x_2, \ldots, x_k)$.

$$s = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - f(x_1, x_2, \ldots, x_k; \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p))^2}{n-p}}$$

The divisor $n - p$ is used because after estimating the p β's from the data, there are only $n - p$ unconstrained residuals left to estimate σ.

Estimated and True Values of σ

Data Set	**True** σ	**Estimated** σ	**95% Confidence Interval for** σ
Calibration	4	3.52	(2.66 , 5.21)
Concrete Strength	173.85	154.75	(127.60 , 196.69)
Plastic Containers	1.7076	1.49	(1.08, 2.39)

s, the estimate of σ from the data and model, is called the residual standard deviation.

Correlation of Parameter Estimates

As we saw earlier, the least squares estimate of the intercept in the straight line model simplifies to

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

which implies that the estimates of the slope and intercept of the line can be linearly related to one another or *correlated*.

The same type of relationship holds for parameters estimates from other more complicated models too.

The correlation between the parameter estimates affects uncertainties of the parameters and has to be taken into account in uncertainty computations.

Variance-Covariance Matrix

A *variance-covariance matrix* summarizes the information needed to compute uncertainties for different quantities derived from the model.

The variance-covariance matrix is a $p \times p$ matrix with the variance of each parameter estimate on the diagonal and the covariance between different pairs of parameters estimates off the diagonal.

The covariance between two parameters estimates is a measure of the strength of the linear relationship between the estimates just like the correlation is, except the covariance is not normalized to lie between -1 and 1.

Variance-Covariance Matrix

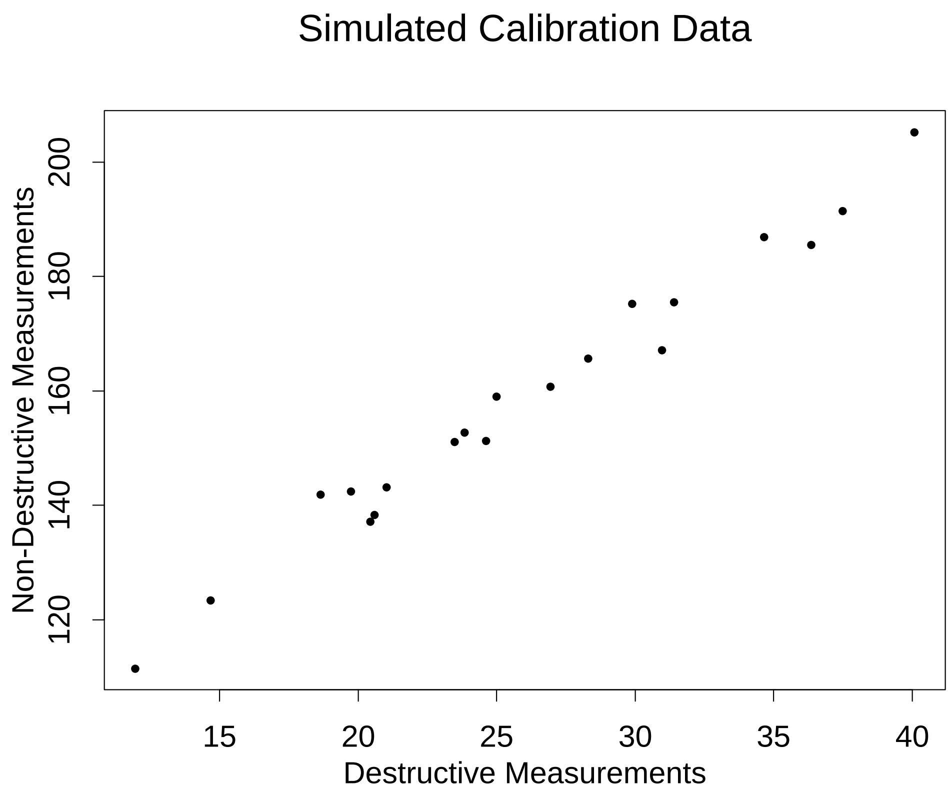
The formula for the variance-covariance matrix is

$$V = s^2(X^T X)^{-1}$$

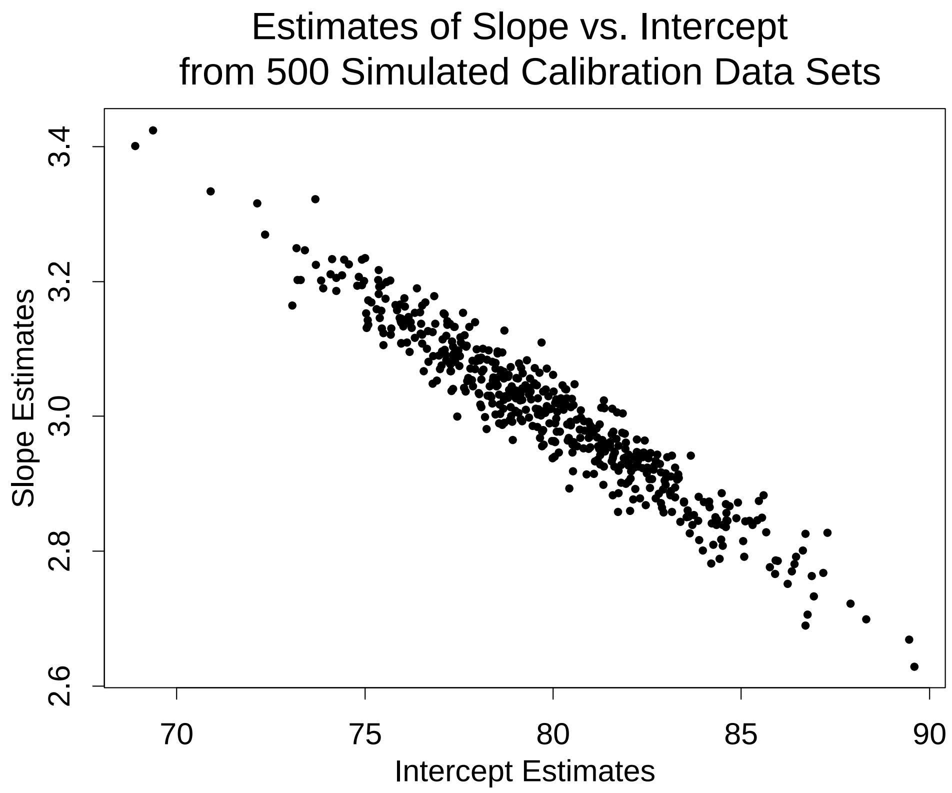
where

- s is the residual standard deviation, and
- X is a matrix with n rows and $p = k + 1$ columns, the first of which is a column of 1's for the intercept term, and the rest of which are each given by the n values of one of the k variables used to fit the model,

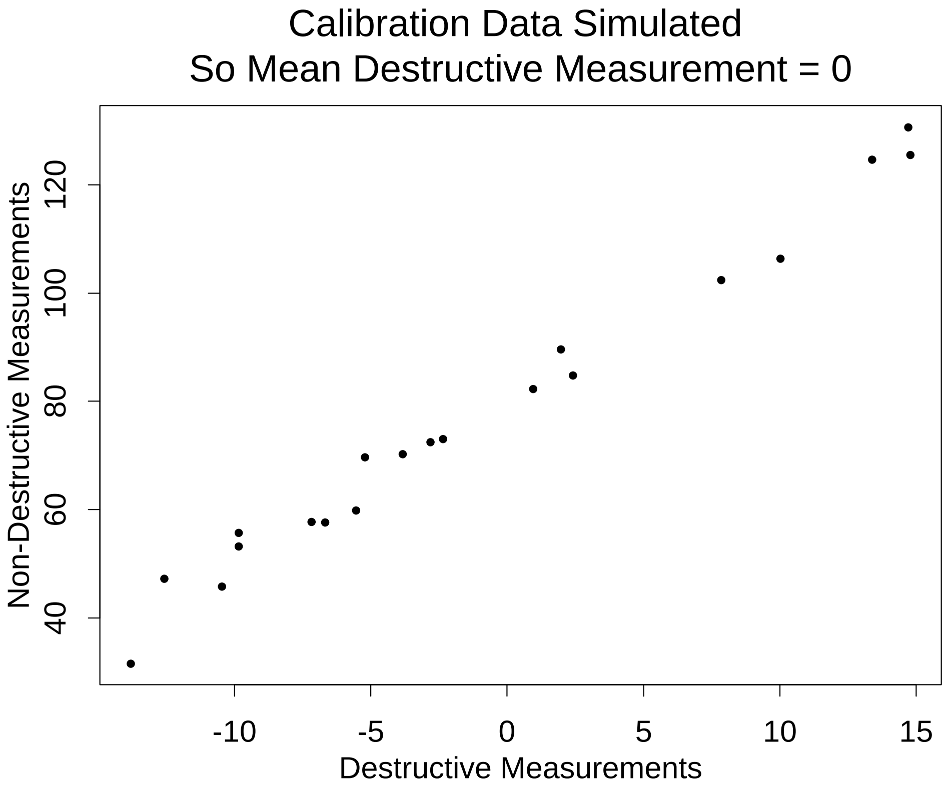
$$X = [1|x_1|x_2|\dots|x_k],$$

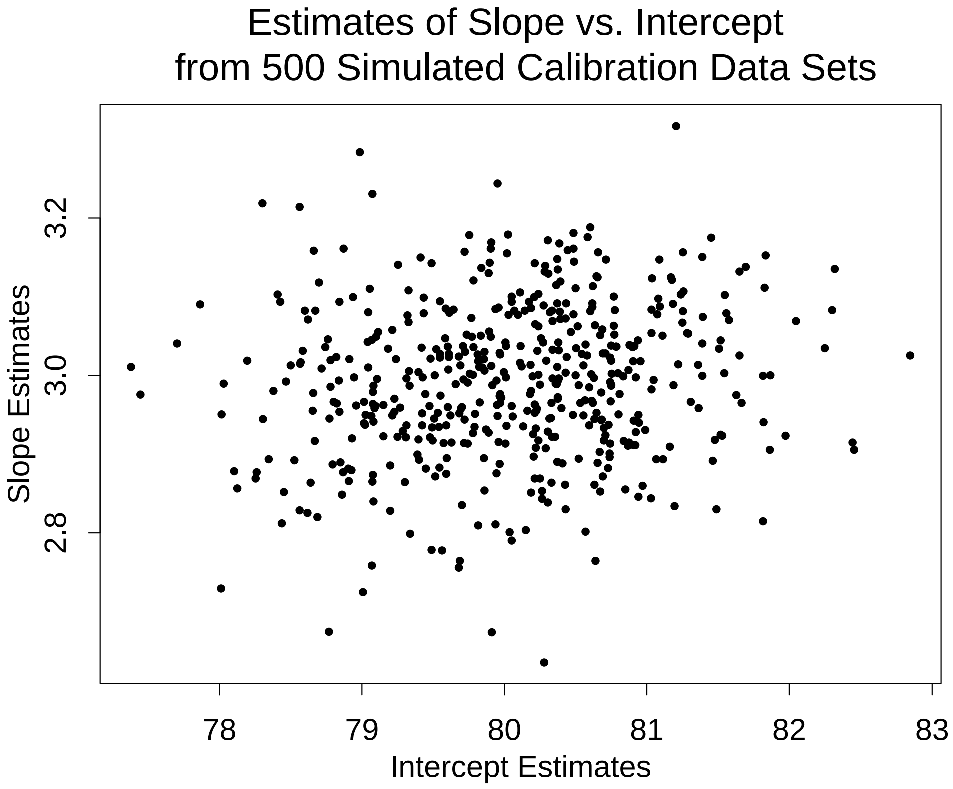


Simulated Calibration Data



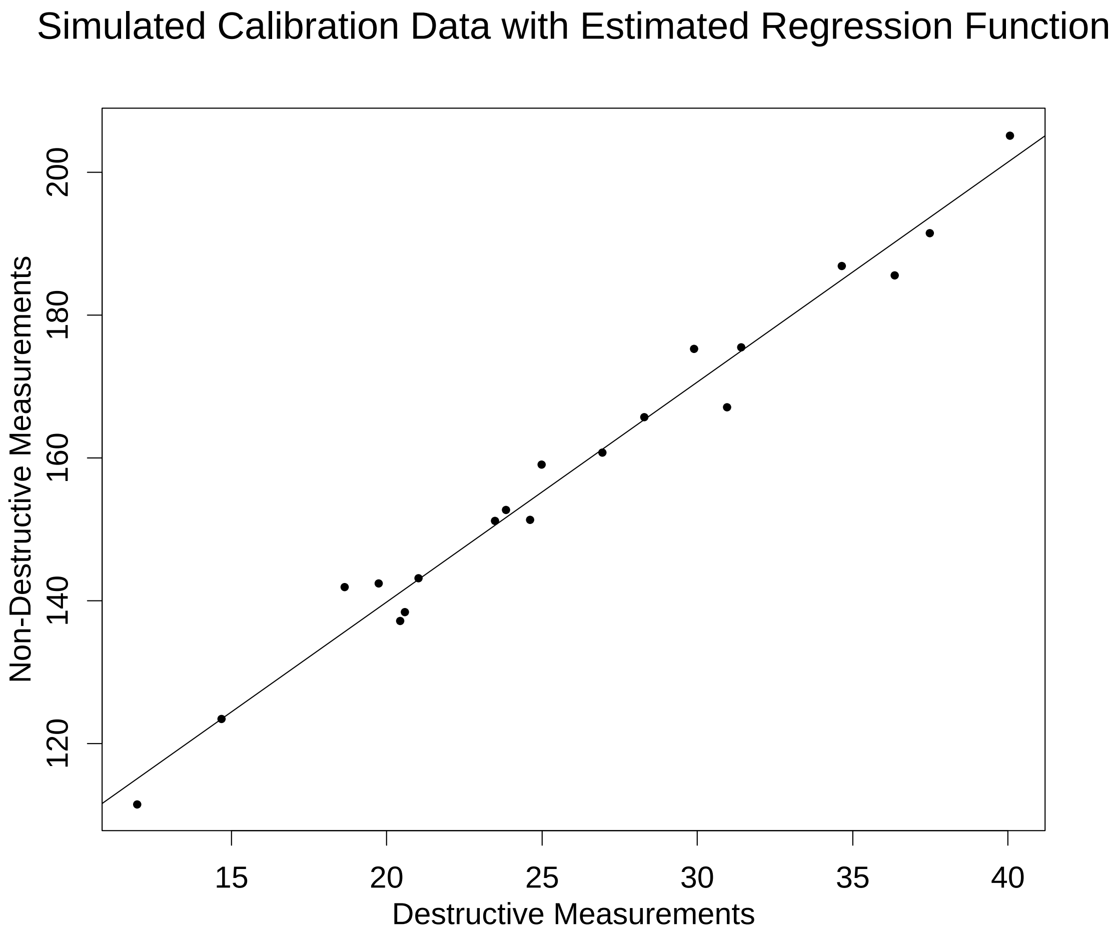
Estimates of Slope vs. Intercept
from 500 Simulated Calibration Data Sets


Calibration Data Simulated
So Mean Destructive Measurement = 0
Non-Destructive Measurements
40
60
80
100
120
-10
-5
0
5
10
15
Destructive Measurements


Estimates of Slope vs. Intercept
from 500 Simulated Calibration Data Sets
Slope Estimates
2.8
3.0
3.2
78
79
80
81
82
83
Intercept Estimates

Regression Output for the Concrete Strength Data

```
Formula: str = (b1 + b2*age)/(1 + b3*age)

Parameters:
       Value     Std. Error  t value
b1   157.461000 116.8220000  1.34787
b2   483.603000  44.2236000 10.93540
b3     0.102985   0.0114841  8.96764

Residual standard error: 154.749 on 42 degrees of freedom

Correlation of Parameter Estimates:
     b1     b2
b2 -0.865
b3 -0.829 0.994
```


Simulated Calibration Data with Estimated Regression Function
Non-Destructive Measurements
120
140
160
180
200
15
20
25
30
35
40
Destructive Measurements

Model Validation

The assumptions made about the data and model are used to check the adequacy of the fit via

1. graphical residual analysis, and
2. statistical tests on the values of the parameters.

Graphical residual analysis is one the most important parts of regression analysis.

Residuals ???

The residuals are essentially estimates of the random errors. They can be analyzed to see if they conform with the assumptions of the regression analysis.

Residuals that conform with the assumptions indicate a good fit, while those that do not indicate the need for changes in the model.

The mathematical definition of the residuals is:

$$e_i = y_i - f(x_{1i}, x_{2i}, \ldots, x_{ki}; \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$$

$$i = 1, 2, \ldots, n$$

Simulated Calibration Data with Estimated Regression Line & Residuals

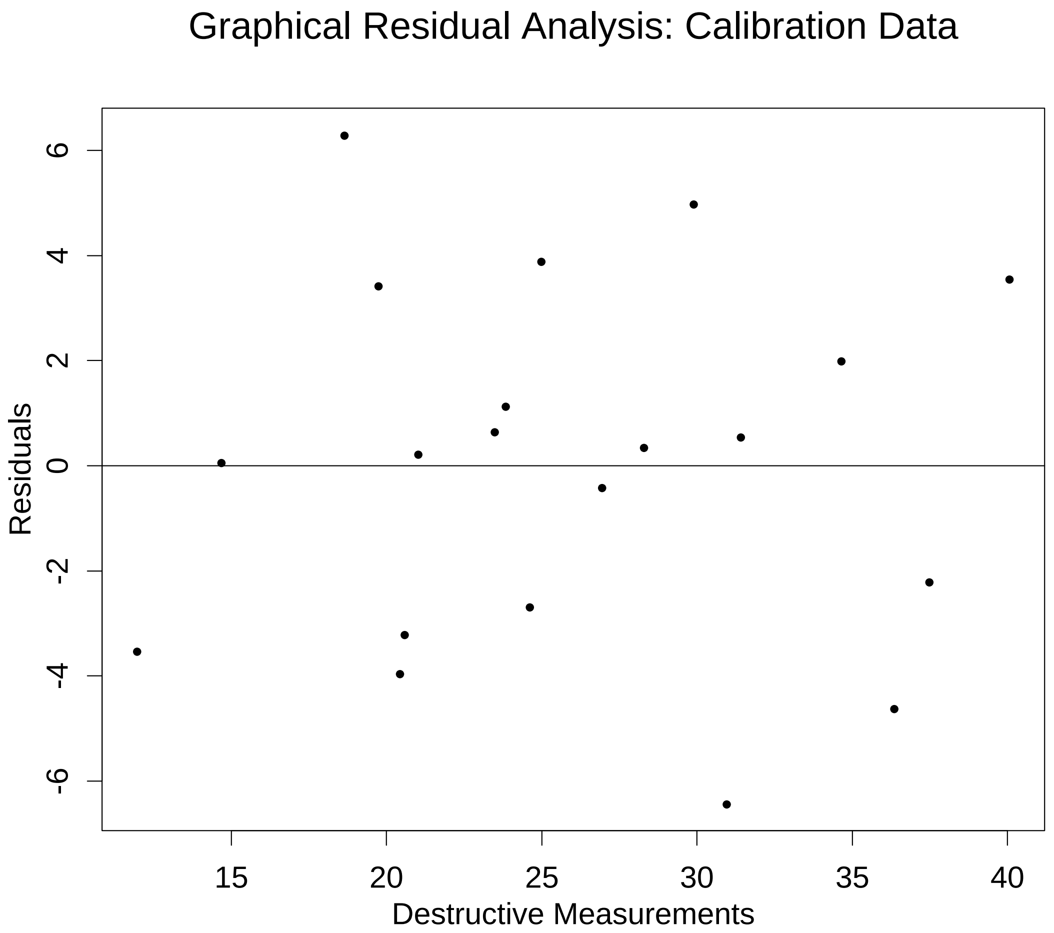
	Destructive	Nondestructive	Predicted	
Run	Measurements	Measurements	Values	Residuals
1	23.83265	152.7652	151.6384	1.12680930
2	24.98922	159.0913	155.2020	3.88922112
3	21.02582	143.2080	142.9900	0.21805514
4	20.58807	138.4274	141.6411	-3.21370688
5	20.43360	137.2064	141.1652	-3.95875318
6	36.34908	185.5820	190.2042	-4.62218645
7	28.29512	165.7351	165.3882	0.34684141
8	18.64261	141.9409	135.6468	6.29413949
9	26.93903	160.7891	161.2098	-0.42067074
10	37.48812	191.5052	193.7138	-2.20866453
11	31.40712	175.5248	174.9770	0.54783253
12	23.48058	151.1957	150.5536	0.64216155
13	14.67612	123.4847	123.4252	0.05956724
14	24.61346	151.3597	154.0443	-2.68460101
15	29.89610	175.3040	170.3212	4.98277061
16	40.06601	205.2072	201.6569	3.55031208
17	11.96256	111.5379	115.0641	-3.52619003
18	30.95920	167.1616	173.5968	-6.43525931
19	19.74407	142.4619	139.0406	3.42126332
20	34.64858	186.9557	184.9646	1.99105836

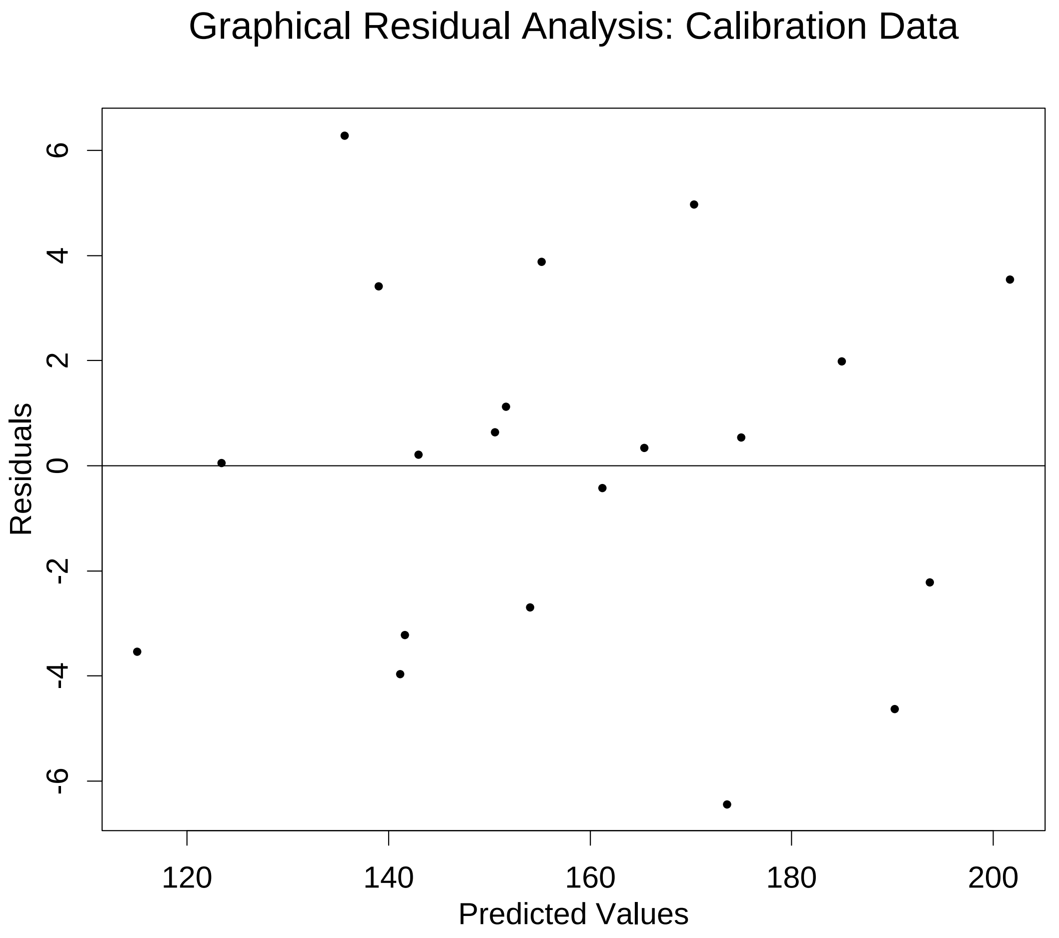
Graphical Residual Analysis

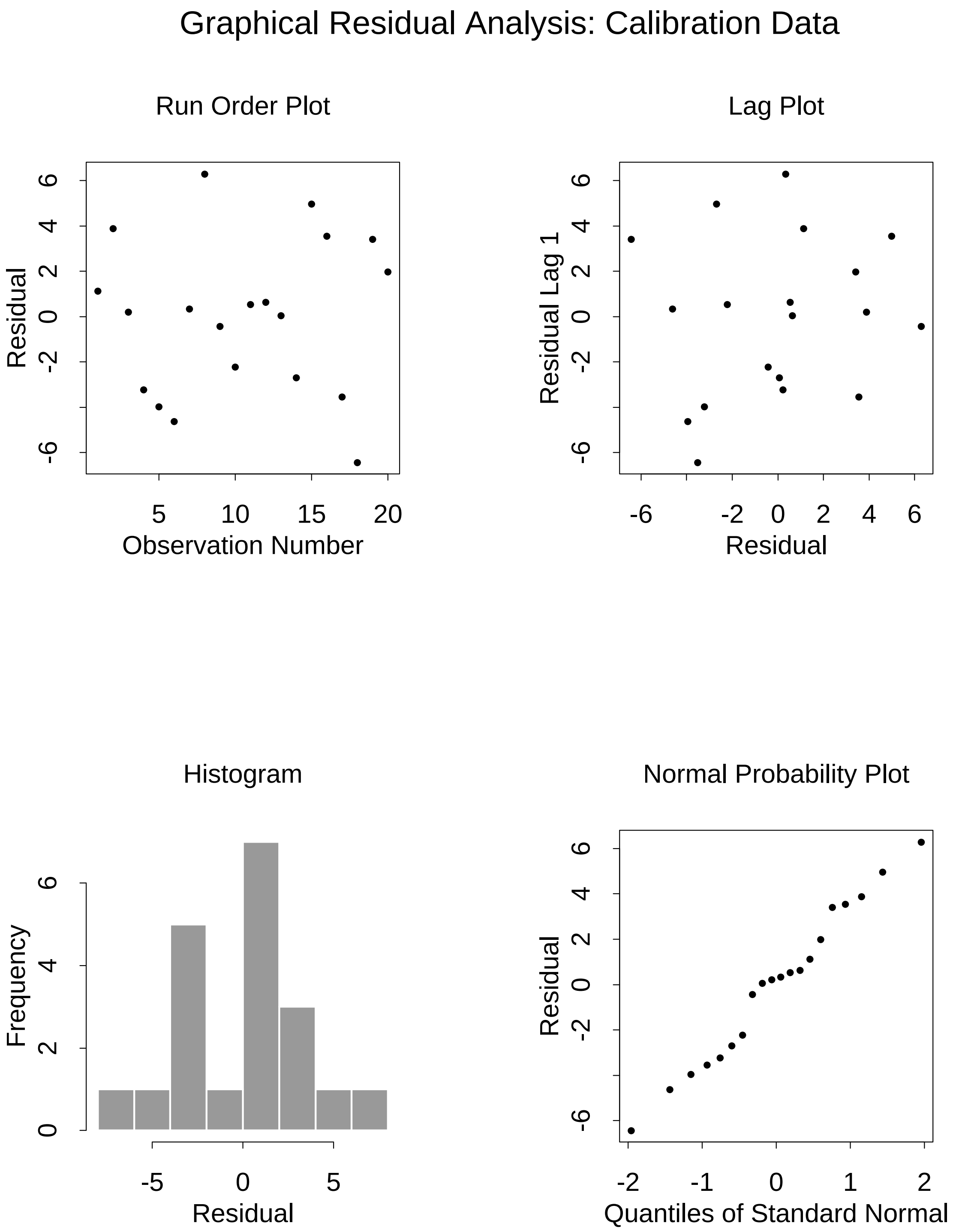
In graphical residual analysis, plots of the residuals are used to to verify that the assumptions underlying reasonable.

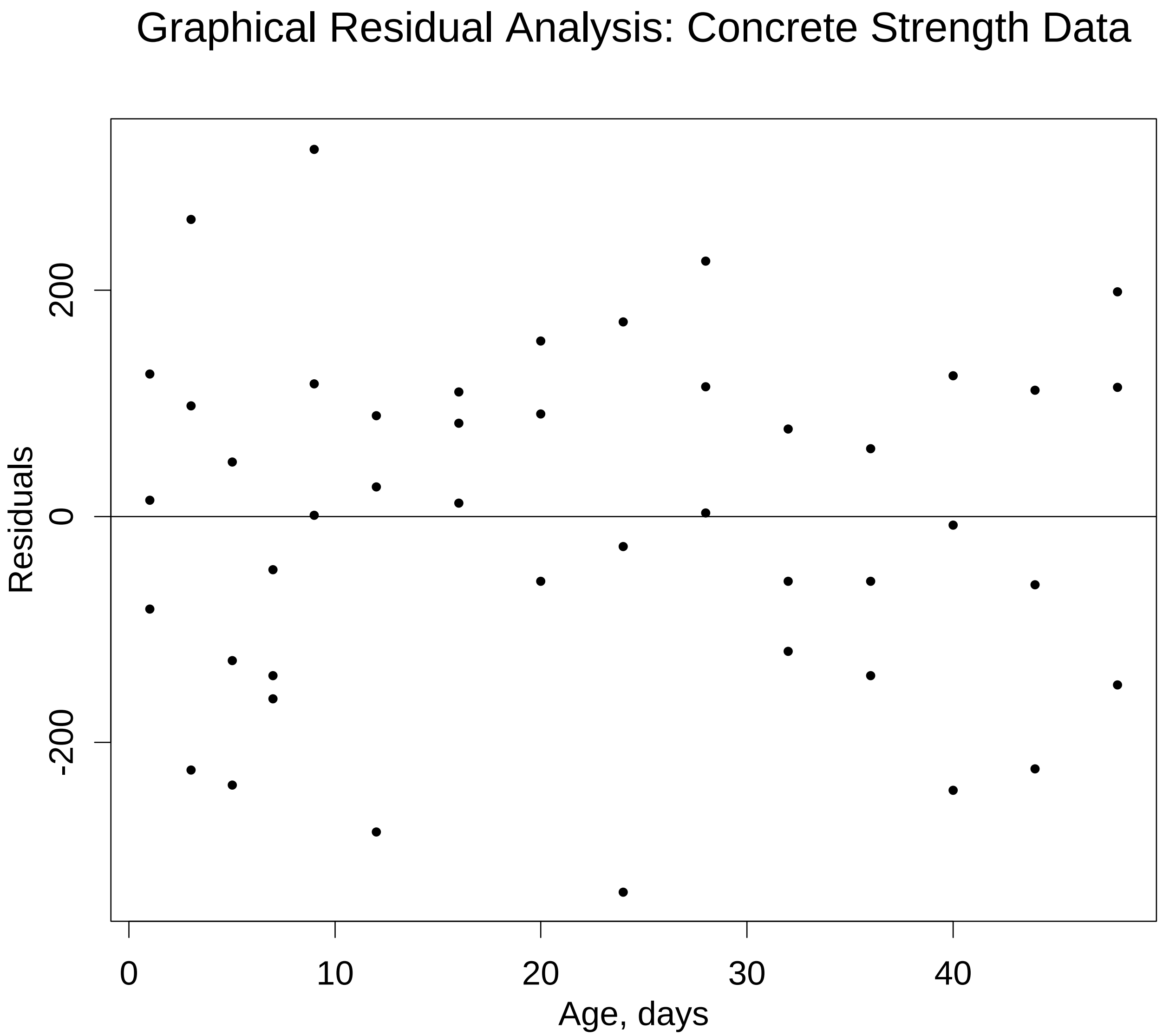
The following plots of the residuals should be made:

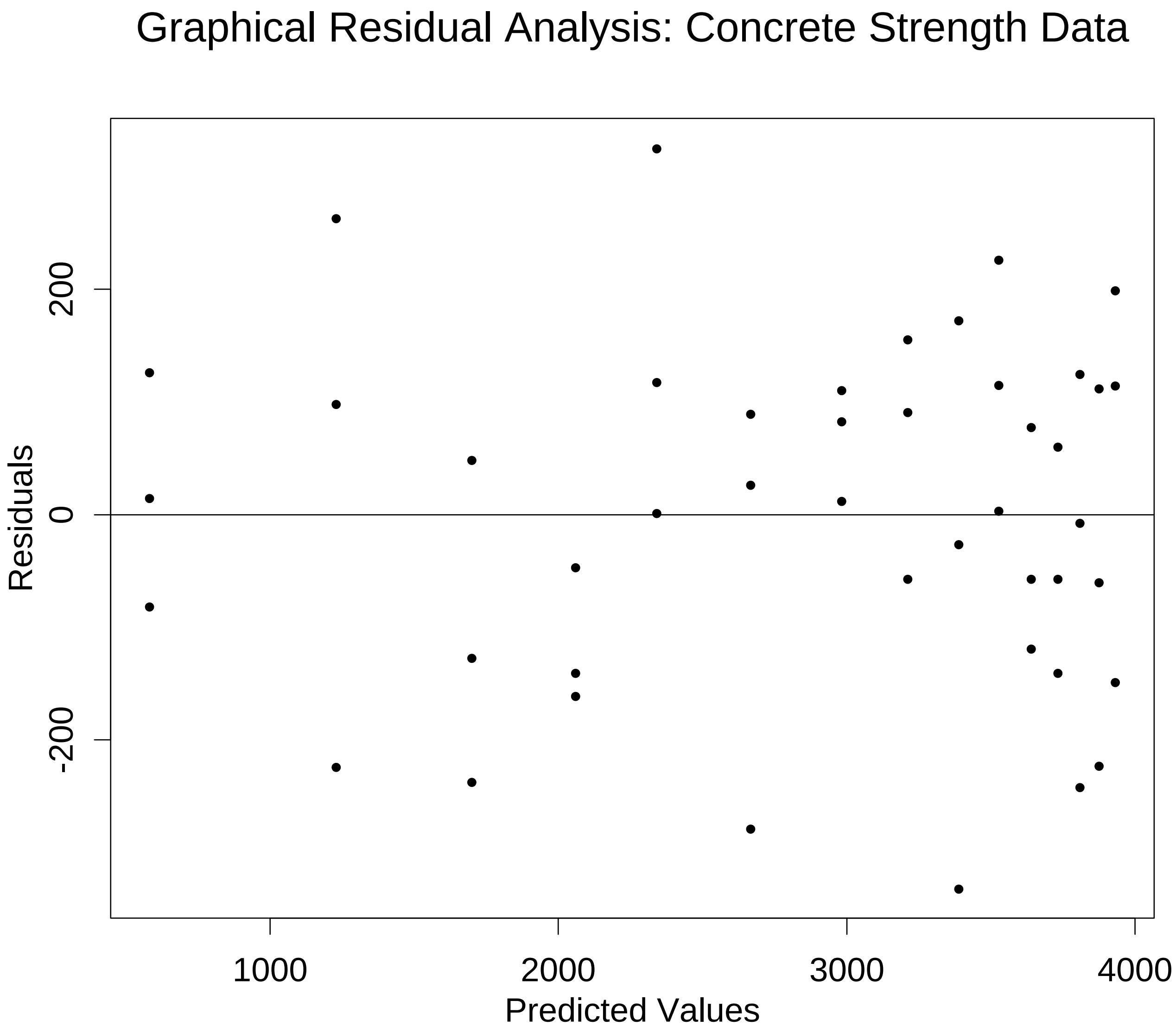
1. e_i vs. all of the predictor variables, x_j
2. e_i vs. the predicted values from the model, $f(x_{1i}, x_{2i}, \ldots, x_{ki}; \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$
3. e_i vs. the order in which the data were collected
4. e_i vs. the lagged residuals, e_{i-1}
5. a histogram of the e_i
6. sorted e_i vs. the quantiles of the standard normal distribution, $\Phi^{-1}(\frac{i-0.375}{n+0.25})$

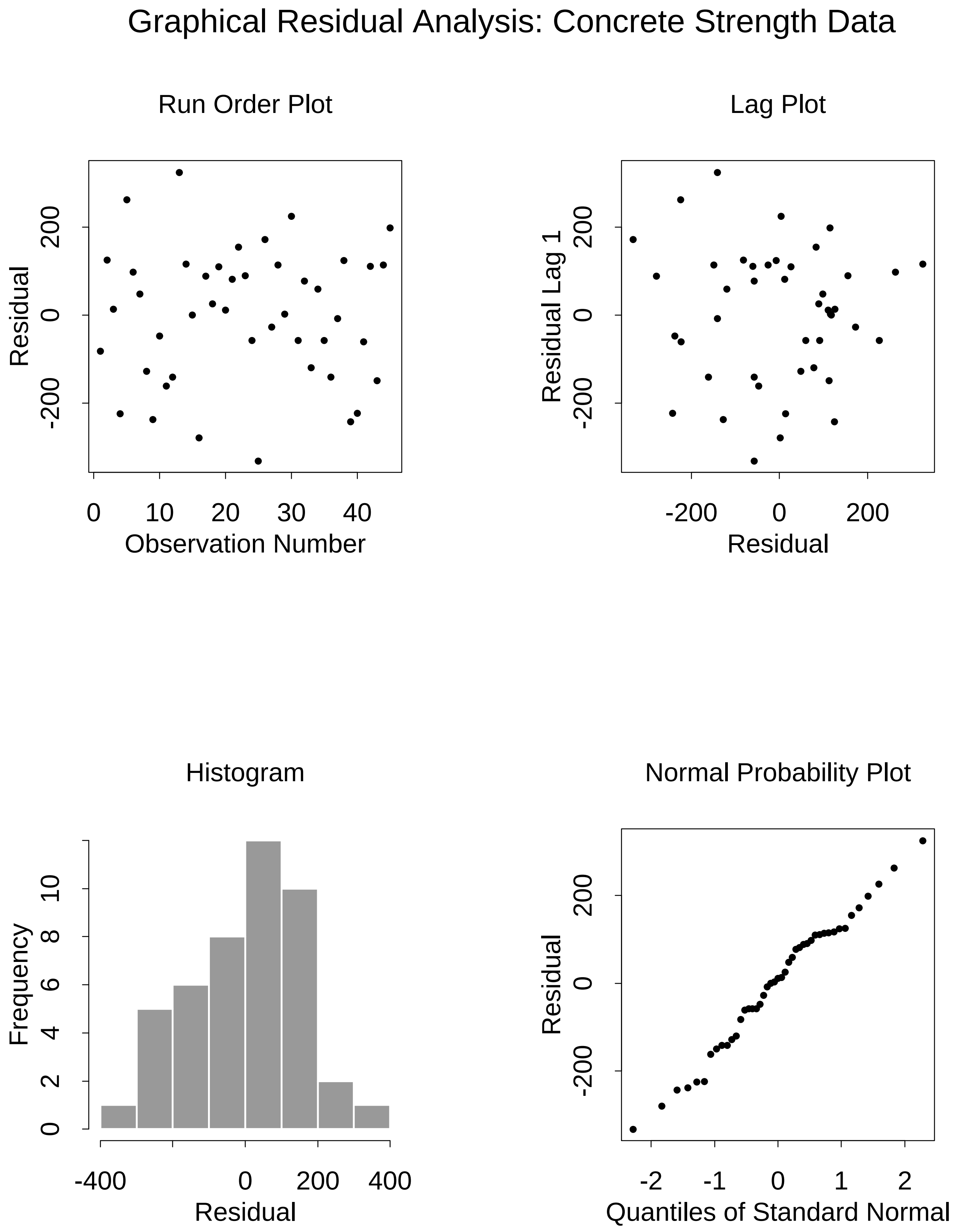

Graphical Residual Analysis: Calibration Data
Residuals
6
4
2
0
-2
-4
-6
15
20
25
30
35
40
Destructive Measurements

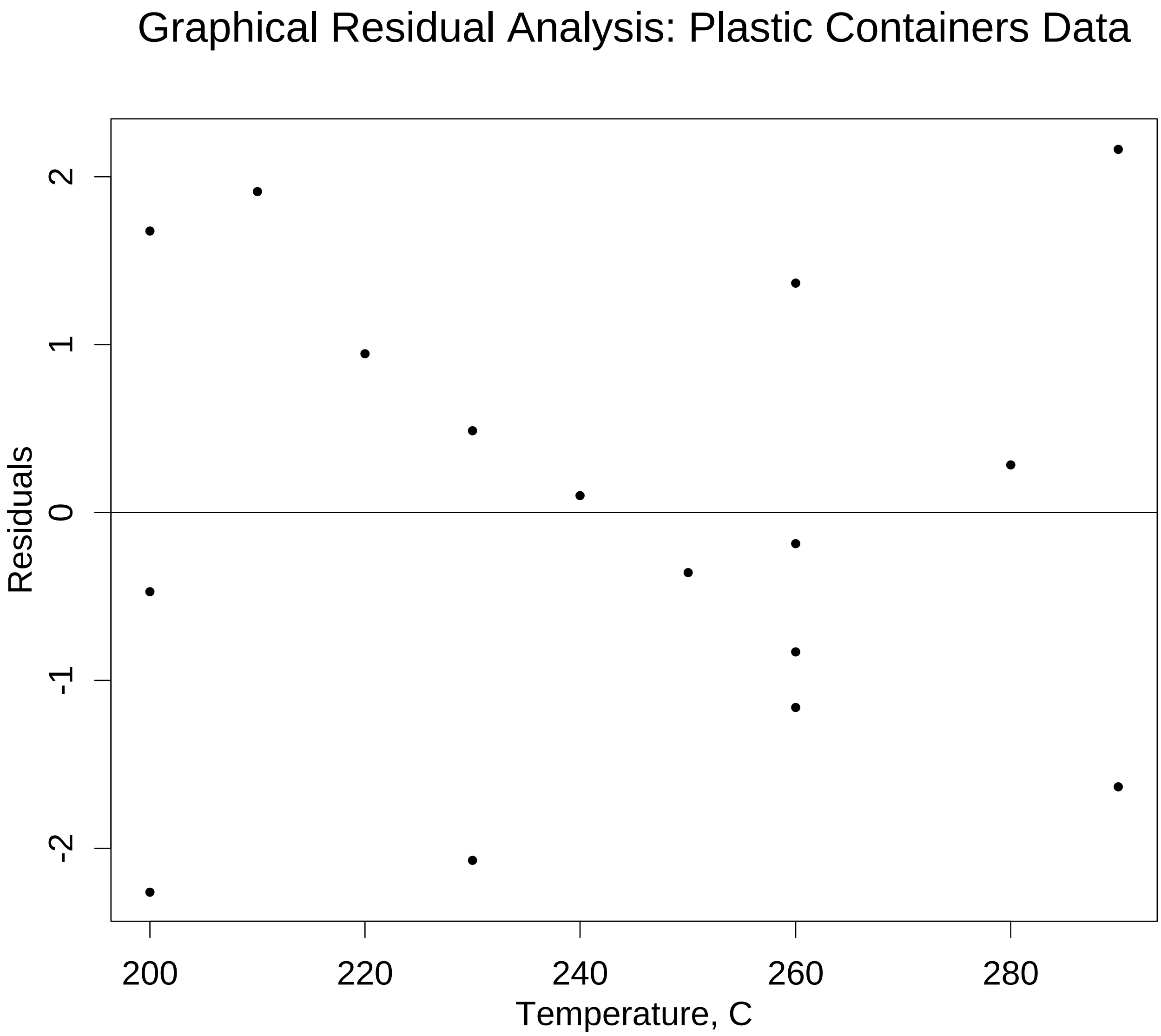

Graphical Residual Analysis: Calibration Data
Residuals
6
4
2
0
-2
-4
-6
120
140
160
180
200
Predicted Values

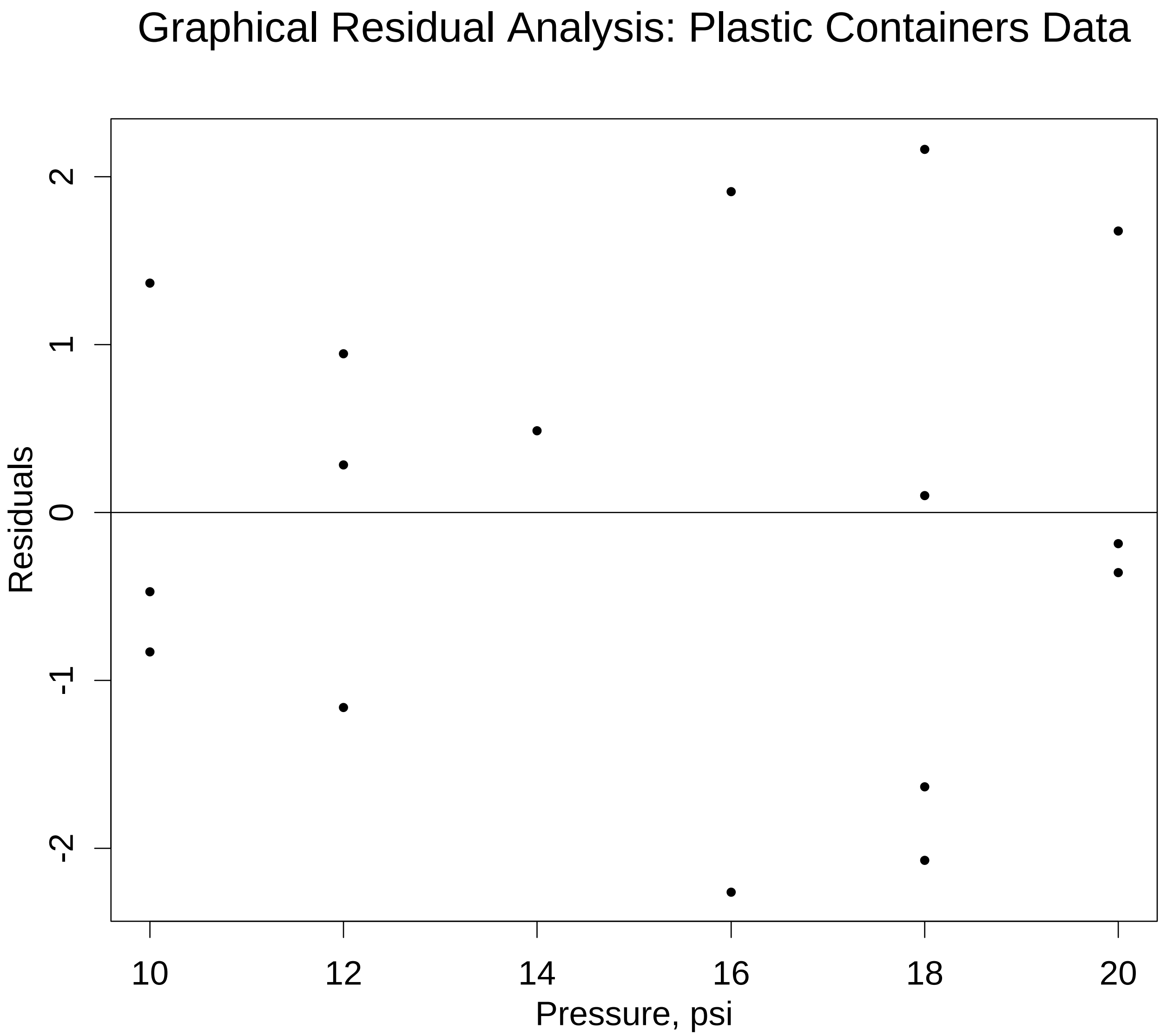

Graphical Residual Analysis: Calibration Data
Run Order Plot
Residual
-6
-2
0
2
4
6
5
10
15
20
Observation Number
Lag Plot
Residual Lag 1
-6
-2
0
2
4
6
-6
-2
0
2
4
6
Residual
Histogram
Frequency
0
2
4
6
-5
0
5
Residual
Normal Probability Plot
Residual
-6
-2
0
2
4
6
-2
-1
0
1
2
Quantiles of Standard Normal

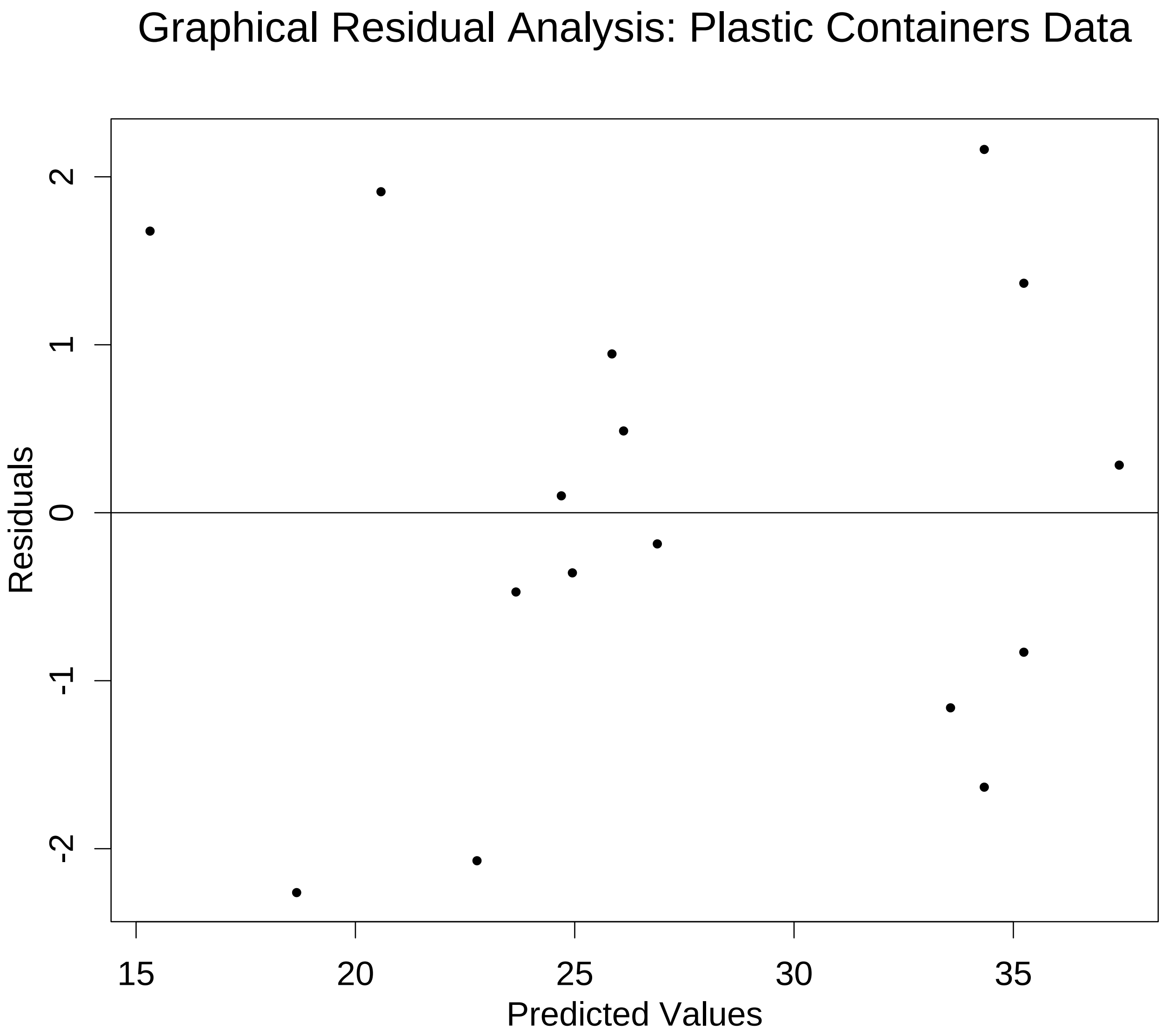

Graphical Residual Analysis: Concrete Strength Data
Residuals
200
0
-200
0
10
20
30
40
Age, days

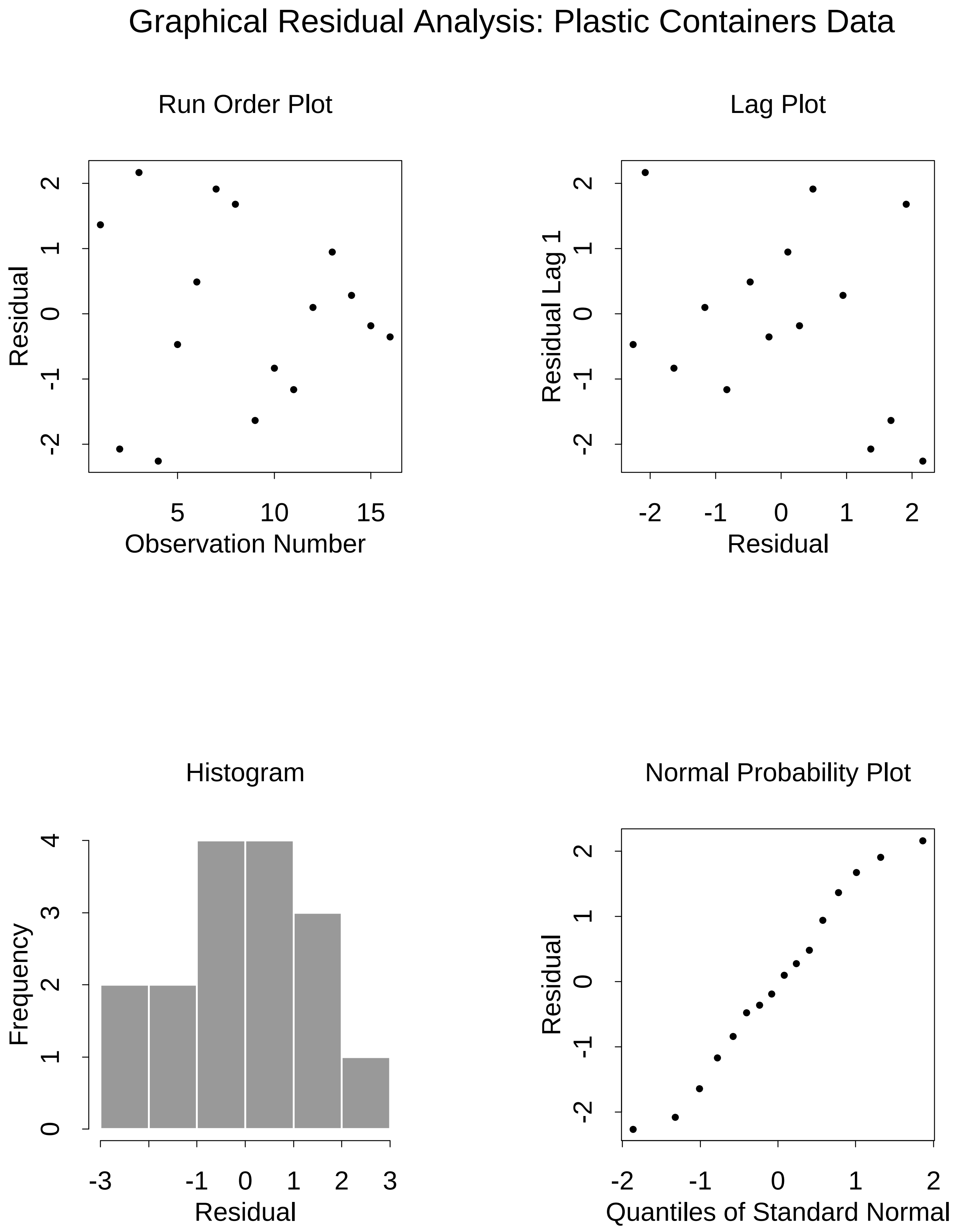

Graphical Residual Analysis: Concrete Strength Data
Residuals
200
0
-200
1000
2000
3000
4000
Predicted Values


Graphical Residual Analysis: Concrete Strength Data
Run Order Plot
Residual
-200
0
200
0
10
20
30
40
Observation Number
Lag Plot
Residual Lag 1
-200
0
200
-200
0
200
Residual
Histogram
Frequency
0
2
4
6
8
10
-400
0
200
400
Residual
Normal Probability Plot
Residual
-200
0
200
-2
-1
0
1
2
Quantiles of Standard Normal


Graphical Residual Analysis: Plastic Containers Data
Residuals
2
1
0
-1
-2
200
220
240
260
280
Temperature, C


Graphical Residual Analysis: Plastic Containers Data
Residuals
2
1
0
-1
-2
10
12
14
16
18
20
Pressure, psi


Graphical Residual Analysis: Plastic Containers Data
Residuals
-2
-1
0
1
2
15
20
25
30
35
Predicted Values


Graphical Residual Analysis: Plastic Containers Data
Run Order Plot
Residual
-2
-1
0
1
2
5
10
15
Observation Number
Lag Plot
Residual Lag 1
-2
-1
0
1
2
-2
-1
0
1
2
Residual
Histogram
Frequency
0
1
2
3
4
-3
-1
0
1
2
3
Residual
Normal Probability Plot
Residual
-2
-1
0
1
2
-2
-1
0
1
2
Quantiles of Standard Normal

F Test for Effectiveness of the Model

H$_0$: All parameters, except β_1 (the intercept) are zero

versus

H$_\mathrm{A}$: At least one parameter, other than β_1, is non-zero

$$F = \frac{\frac{\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{p-1}}{\frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}}$$

$F > \mathrm{F}_{1-\alpha,p-1,n-p} \Rightarrow$ H$_0$ should be rejected

Tests on the Individual Parameters

Test of $\mathrm{H_0}$: $\beta_j = 0$ vs. $\mathrm{H_A}$: $\beta_j \neq 0$

$$T = \frac{(\hat{\beta}_j - 0)}{SD(\hat{\beta}_j)}$$

$|T| > \mathrm{t}_{1-\alpha/2,n-p} \Rightarrow \mathrm{H_0}$ should be rejected

$SD(\hat{\beta}_j)$ is the standard deviation of the parameter estimate, which most software will provide automatically.

$SD(\hat{\beta}_j)$ is a function of s which accounts for the random variation in the data, the amount of averaging inherent in the computation of the estimate, and the covariances between the estimated parameters in the model.

Regression Output for the Calibration Data

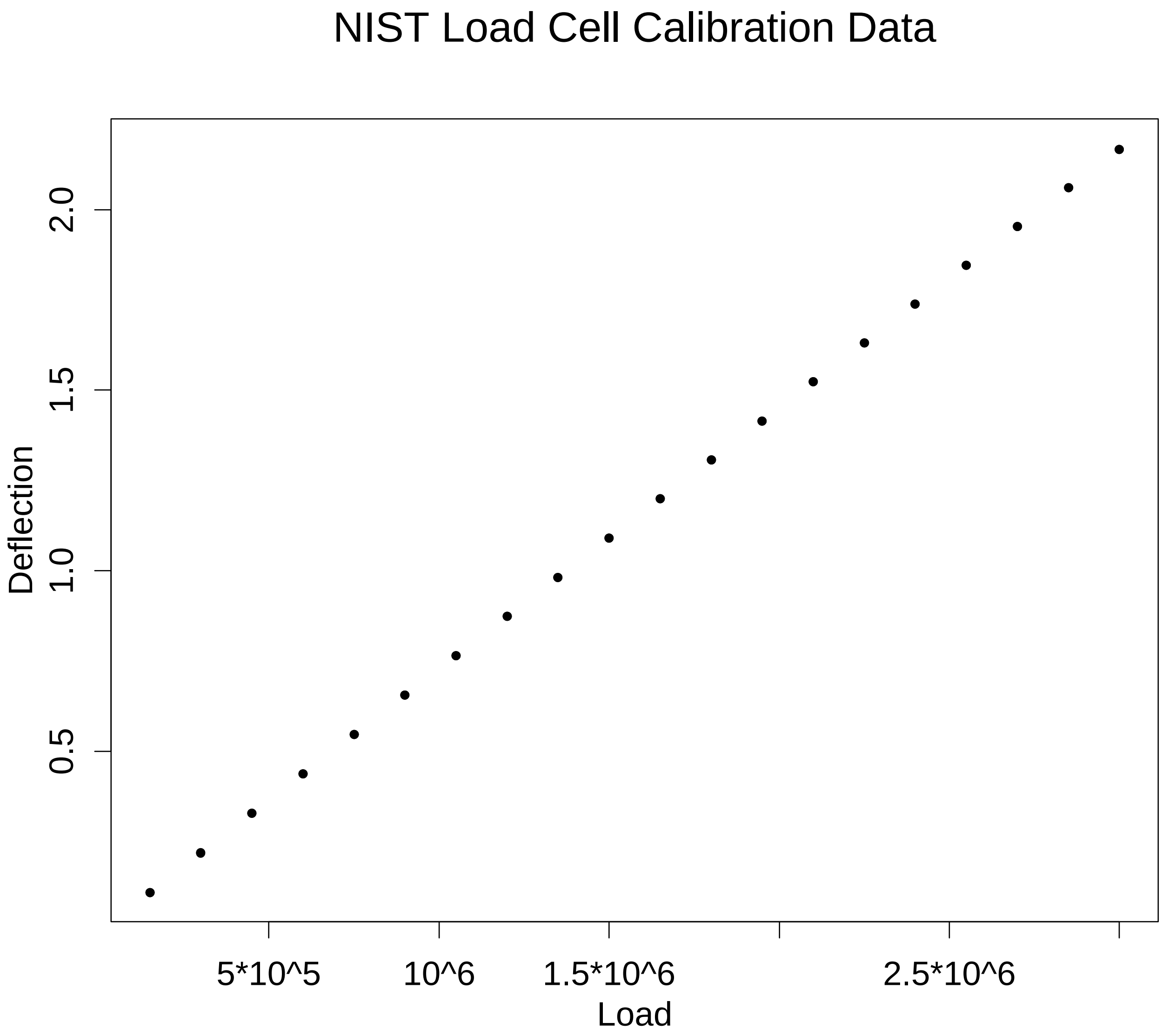
```
N=20

Residual Standard Error = 3.5235

Multiple R-Square = 0.9791

F-statistic = 844.5767 on 1 and 18 df, p-value = 0

                   coef std.err  t.stat p.value
Intercept       78.2049  2.8672 27.2758       0
Destructive MM   3.0812  0.1060 29.0616       0
```

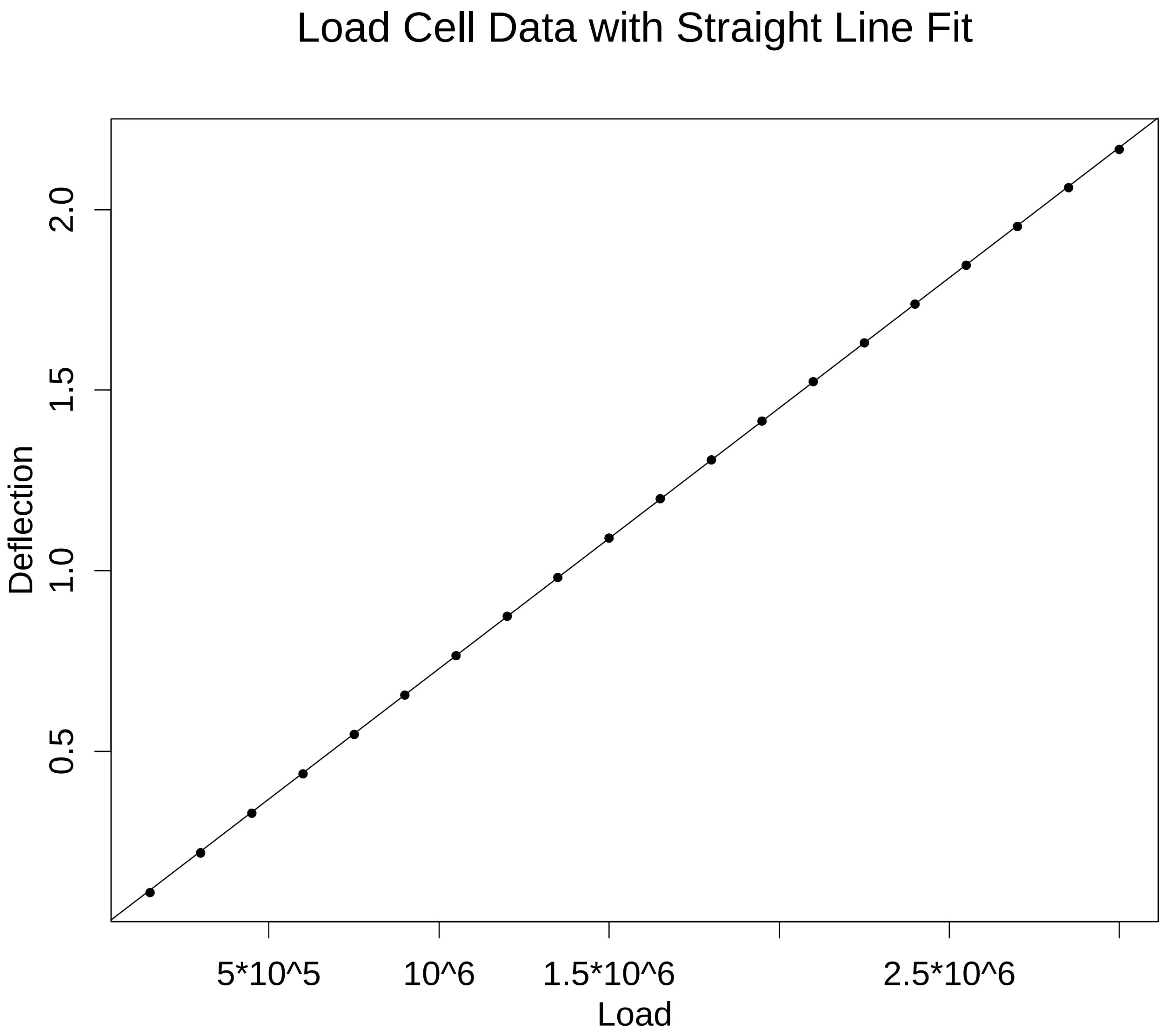
Regression Output for the Concrete Strength Data

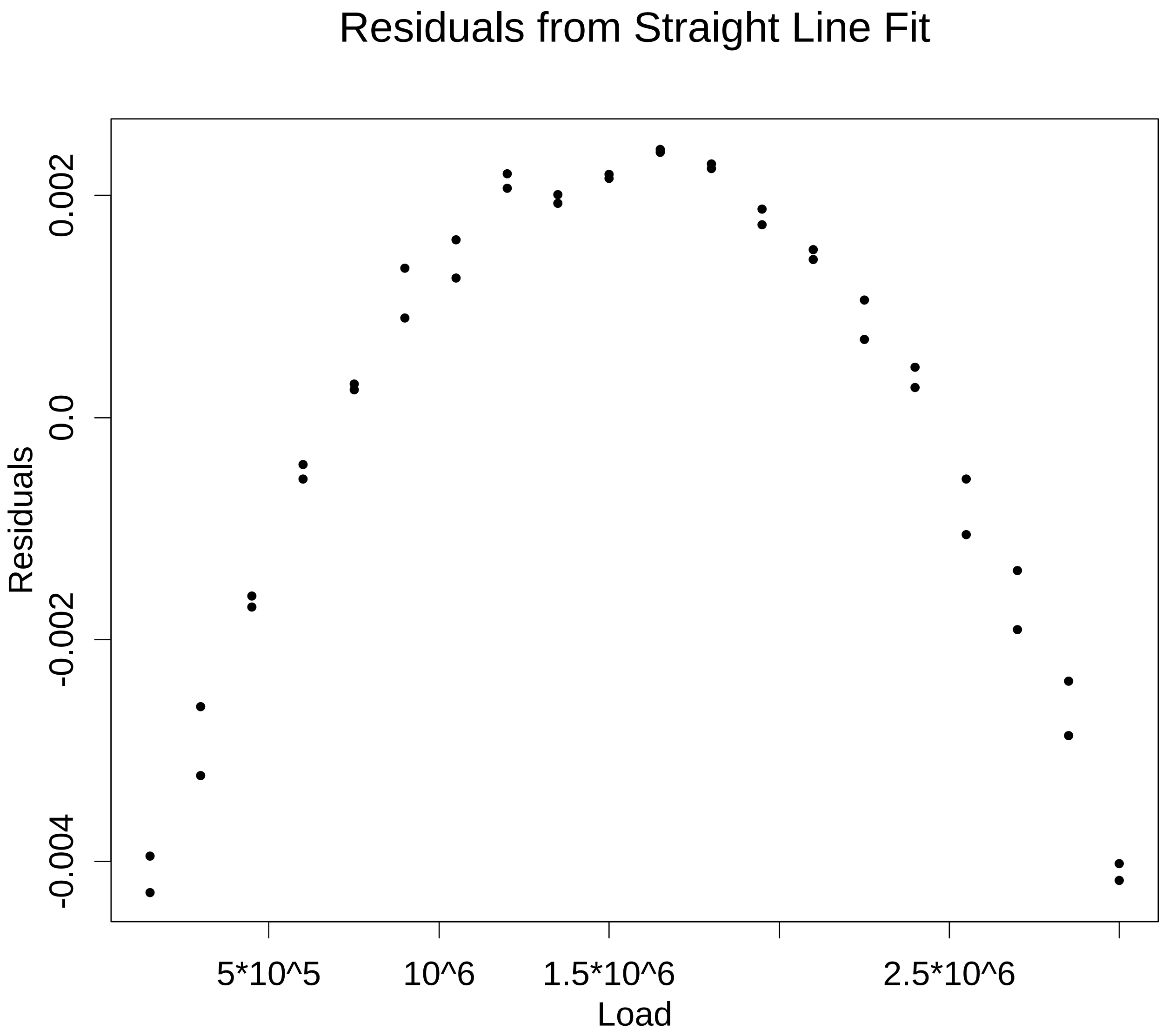
```
Formula: str = (b1 + b2*age)/(1 + b3*age)

Parameters:
       Value      Std. Error  t value
b1   157.461000 116.8220000  1.34787
b2   483.603000  44.2236000 10.93540
b3     0.102985   0.0114841  8.96764

Residual standard error: 154.749 on 42 degrees of freedom

Correlation of Parameter Estimates:
     b1     b2
b2 -0.865
b3 -0.829 0.994
```

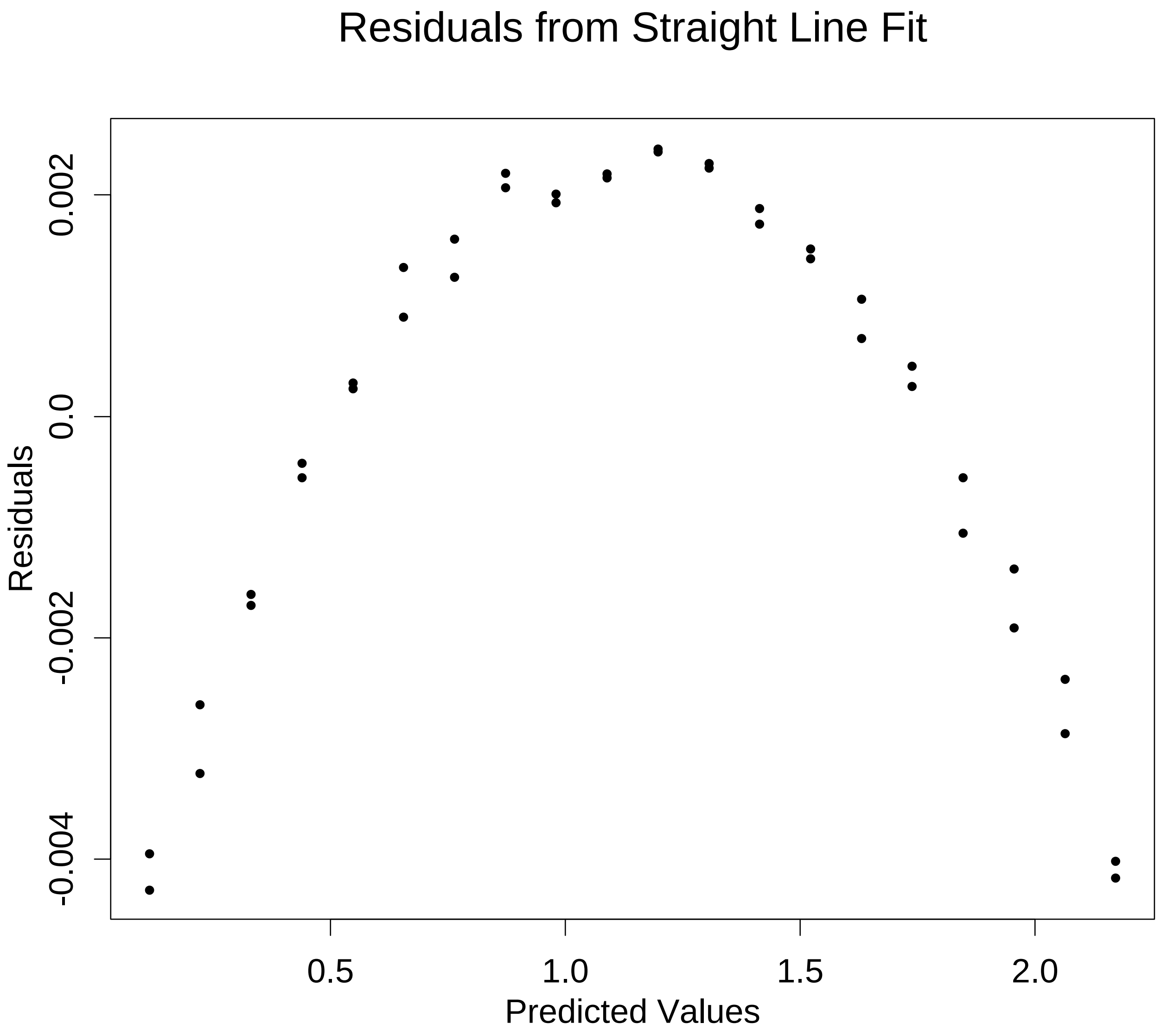
Regression Output for the Plastic Containers Data

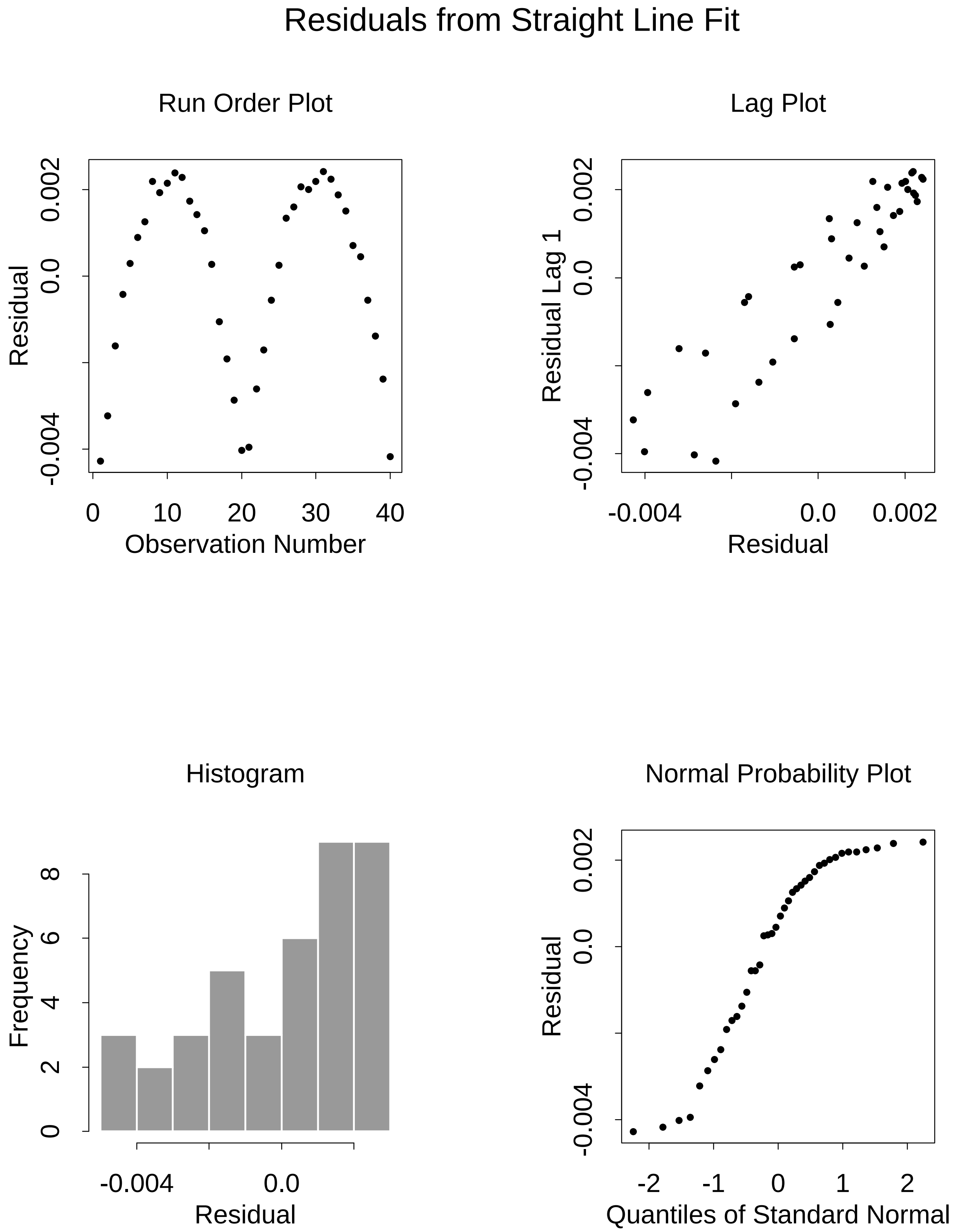
```
N=16

Residual Standard Error = 1.4874

Multiple R-Square = 0.9593

F-statistic = 153.3614 on 2 and 13 df, p-value = 0

               coef  std.err    t.stat  p.value
Intercept   -6.5224   3.3689   -1.9361   0.0749
Pressure    -0.8348   0.1015   -8.2283   0.0000
Temperature  0.1927   0.0124   15.5978   0.0000
```


NIST Load Cell Calibration Data
Deflection
0.5
1.0
1.5
2.0
5*10^5
10^6
1.5*10^6
2.5*10^6
Load


Load Cell Data with Straight Line Fit
Deflection
0.5
1.0
1.5
2.0
5*10^5
10^6
1.5*10^6
2.5*10^6
Load

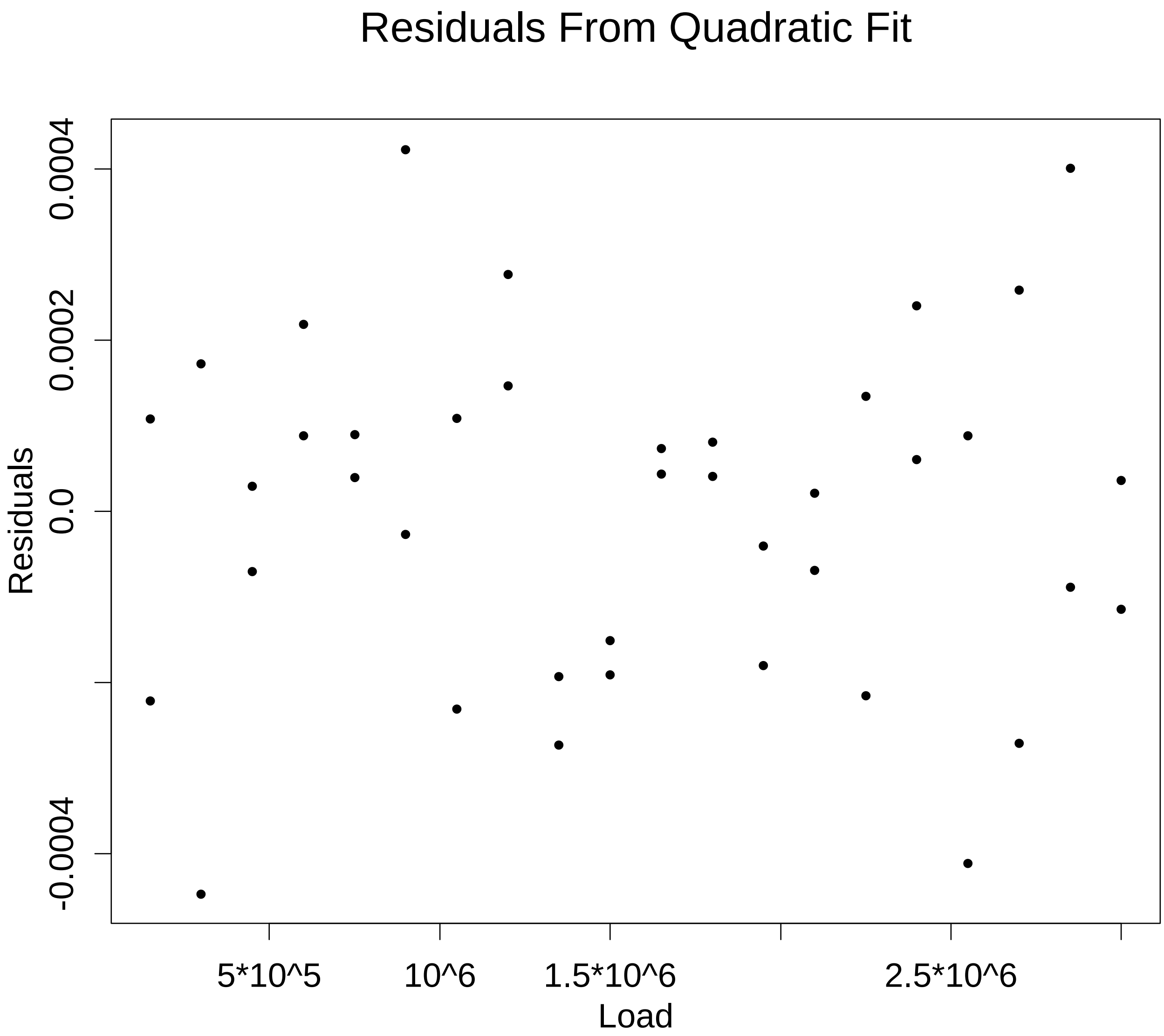

Residuals from Straight Line Fit
Residuals
0.002
0.0
-0.002
-0.004
5*10^5
10^6
1.5*10^6
2.5*10^6
Load

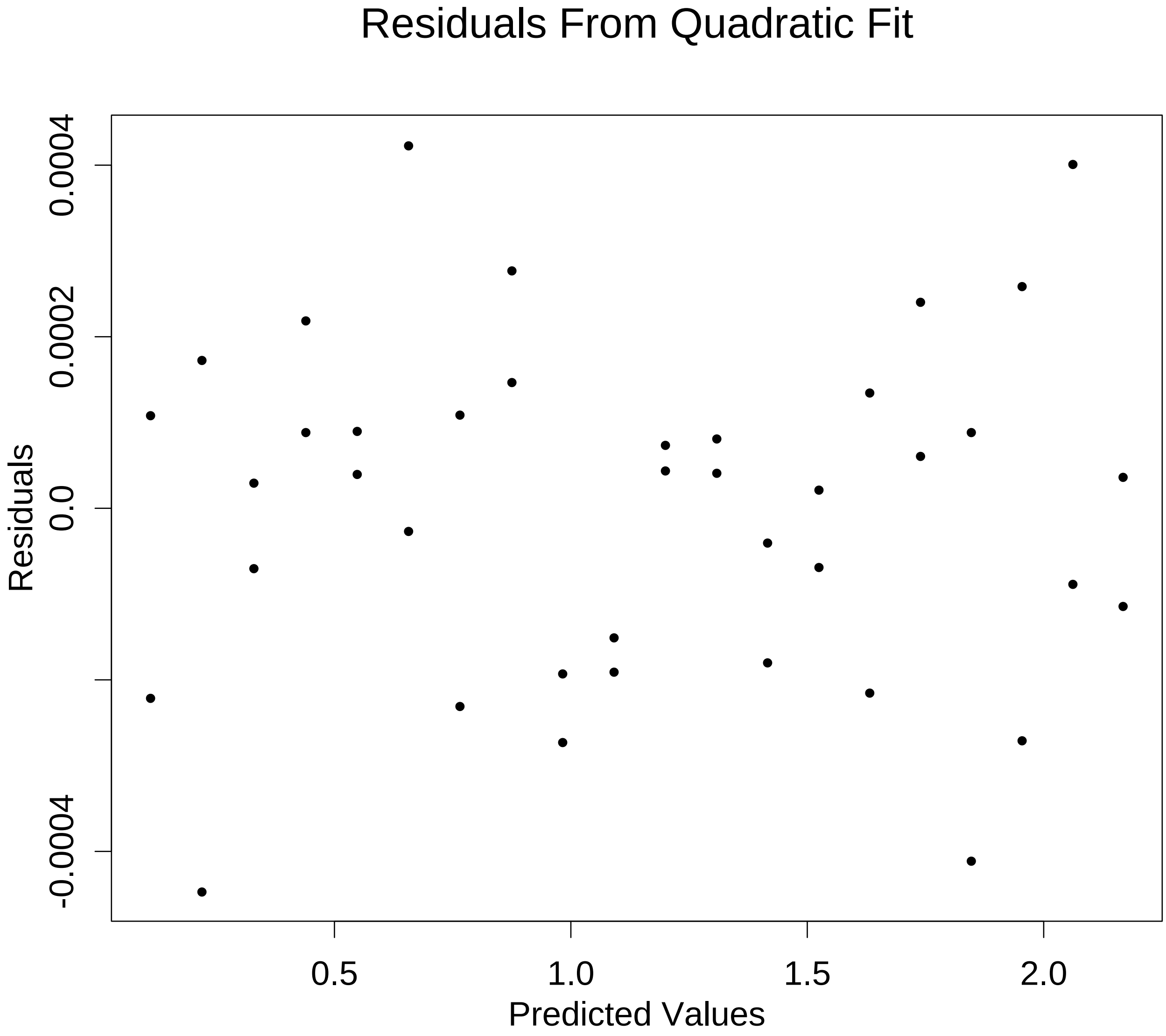

Residuals from Straight Line Fit
Residuals
0.002
0.0
-0.002
-0.004
Predicted Values
0.5
1.0
1.5
2.0

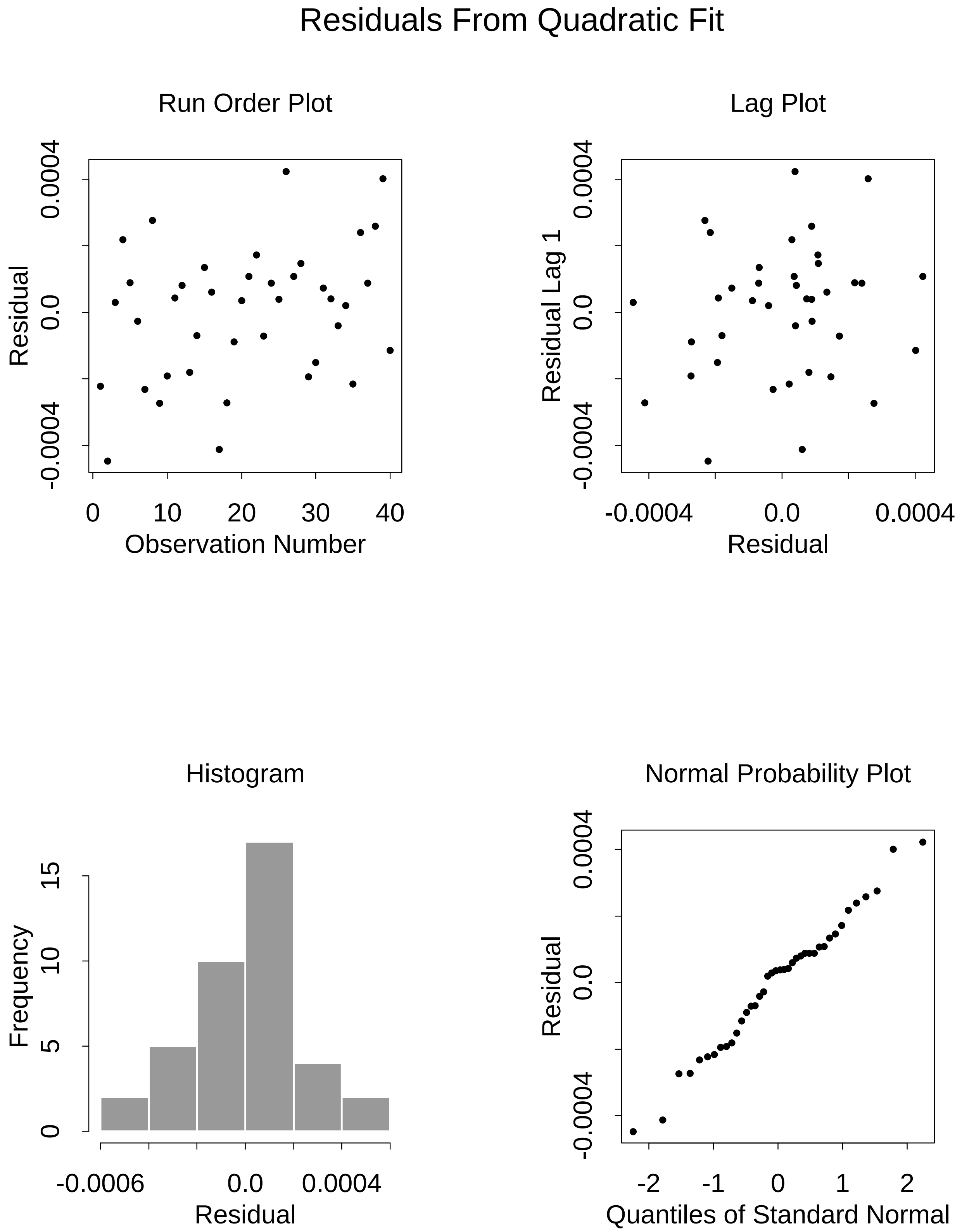

Residuals from Straight Line Fit
Run Order Plot
Residual
0.002
0.0
-0.004
0
10
20
30
40
Observation Number
Lag Plot
Residual Lag 1
0.002
0.0
-0.004
-0.004
0.0
0.002
Residual
Histogram
Frequency
8
6
4
2
0
-0.004
0.0
Residual
Normal Probability Plot
Residual
0.002
0.0
-0.004
-2
-1
0
1
2
Quantiles of Standard Normal

Load Cell Data Regression Output
Straight Line Model

```
N = 40

Residual Standard Error = 0.002171273

Multiple R-Square = 0.9999885

F-statistic = 3309811 on 1 and 38 df, p-value = 0

                  coef      std.err      t.stat      p.value
Intercept 6.149684e-03 7.132052e-04    8.622602 1.772154e-10
Load      7.221026e-07 3.969148e-10 1819.288717 0.000000e+00
```

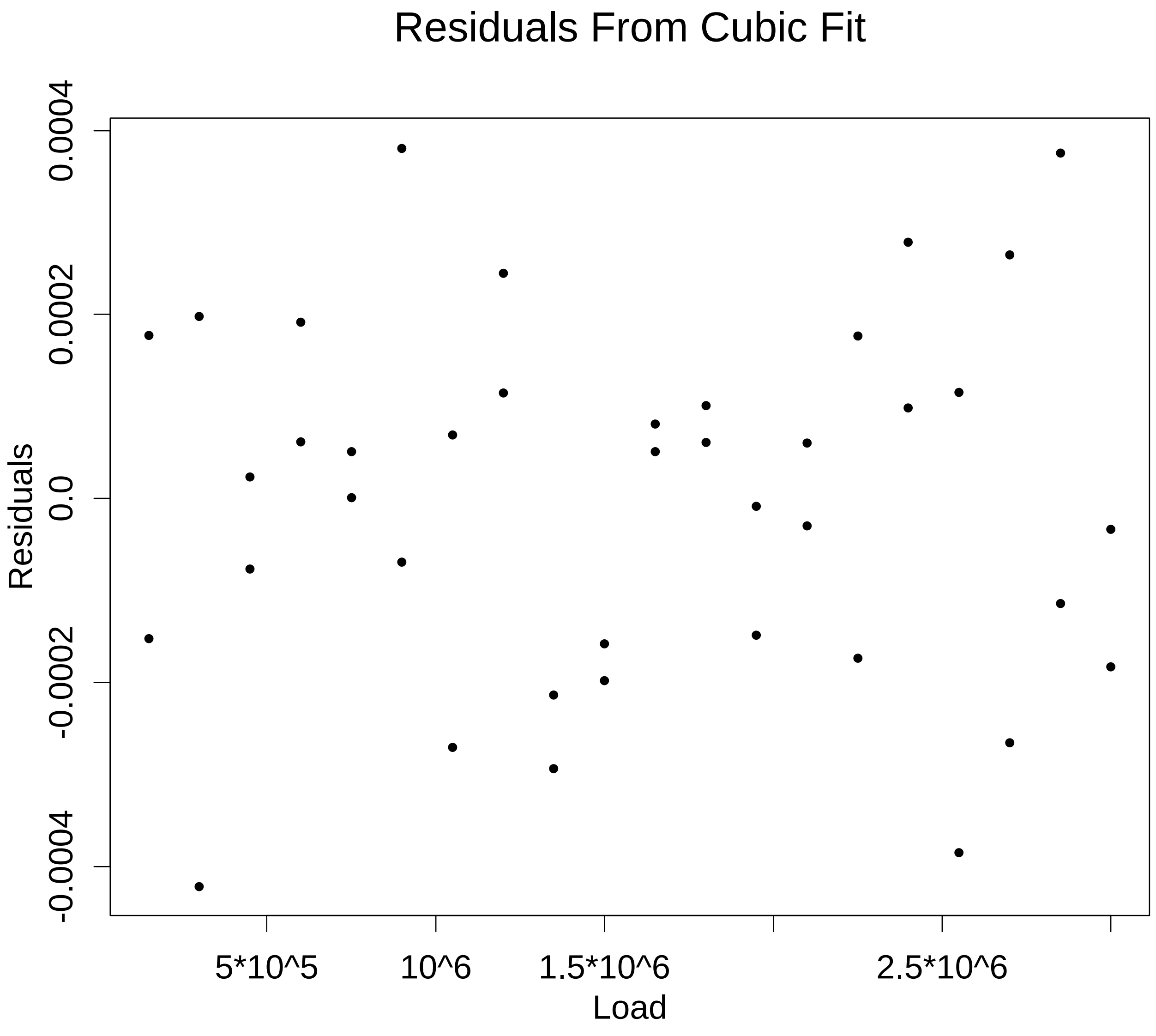

Residuals From Quadratic Fit
Residuals
0.0004
0.0002
0.0
-0.0004
5*10^5
10^6
1.5*10^6
2.5*10^6
Load

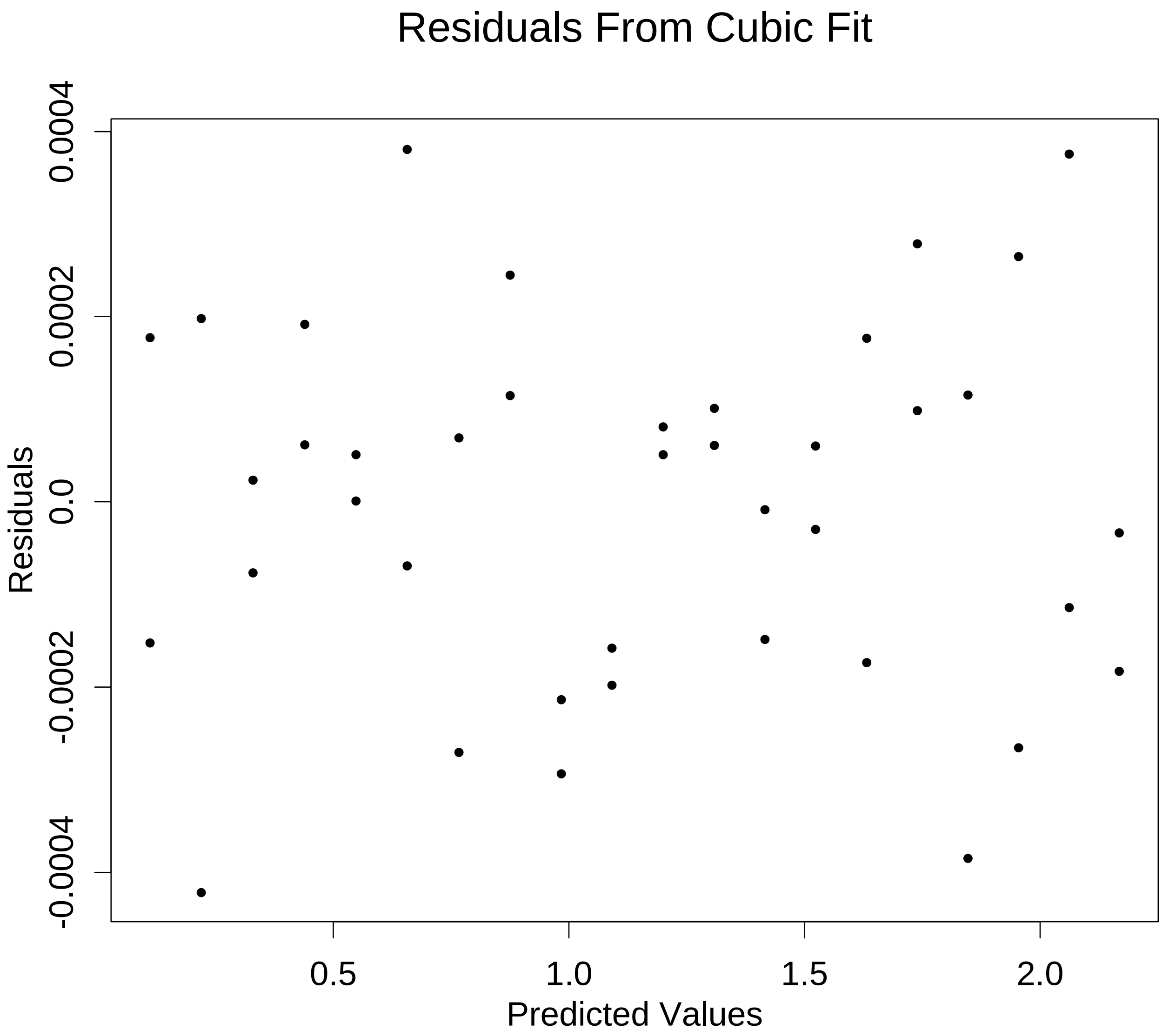

Residuals From Quadratic Fit
Residuals
0.0004
0.0002
0.0
-0.0004
0.5
1.0
1.5
2.0
Predicted Values

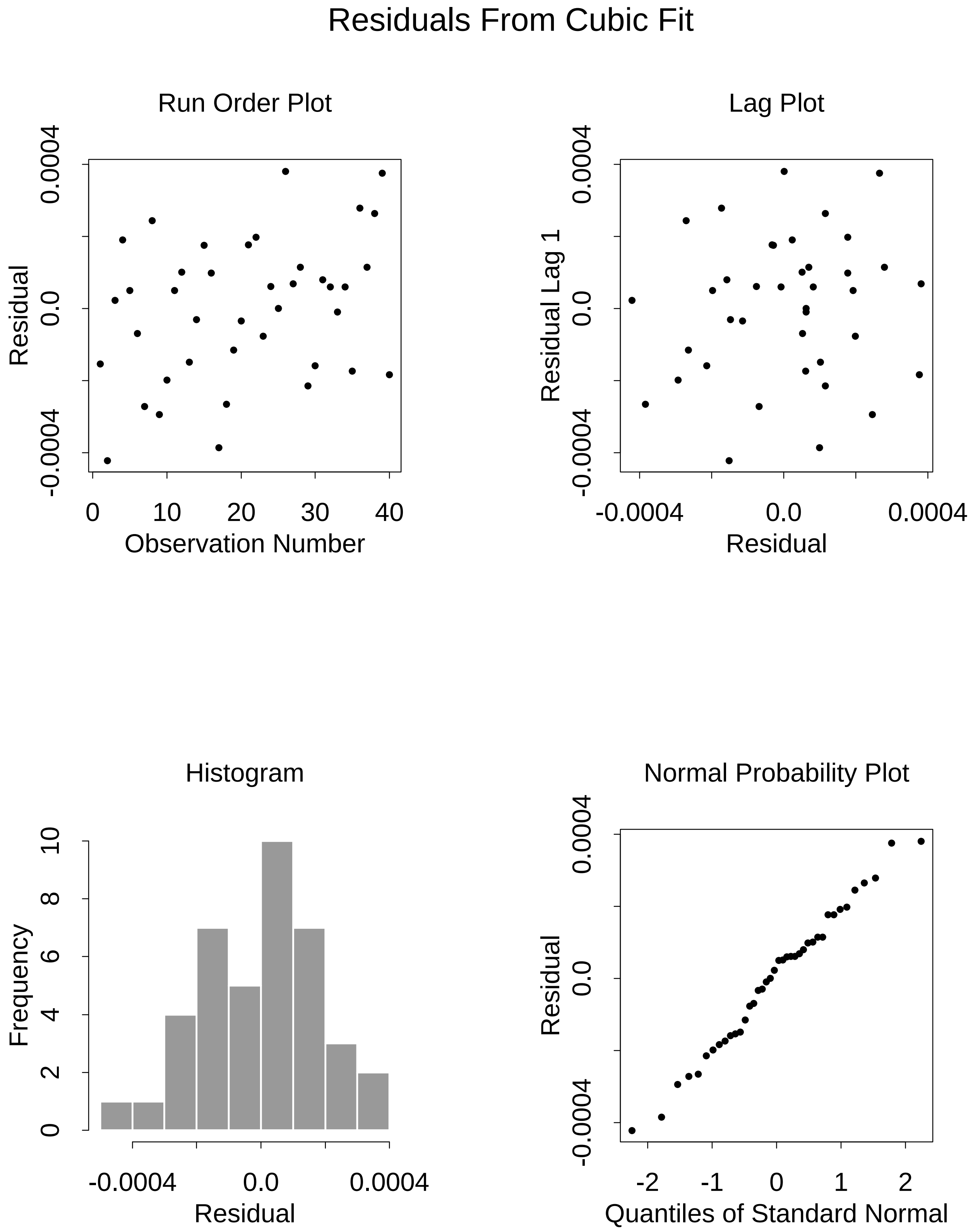

Residuals From Quadratic Fit
Run Order Plot
Residual
-0.0004
0.0
0.0004
0
10
20
30
40
Observation Number
Lag Plot
Residual Lag 1
-0.0004
0.0
0.0004
-0.0004
0.0
0.0004
Residual
Histogram
Frequency
0
5
10
15
-0.0006
0.0
0.0004
Residual
Normal Probability Plot
Residual
-0.0004
0.0
0.0004
-2
-1
0
1
2
Quantiles of Standard Normal

Load Cell Data Regression Output
Quadratic Model

```
N = 40

Residual Standard Error = 0.0002051774

Multiple R-Square = 0.9999999

F-statistic = 185330866 on 2 and 37 df, p-value = 0

                   coef      std.err       t.stat      p.value
Intercept  6.735658e-04 1.079386e-04     6.240267 2.970542e-07
Load       7.320592e-07 1.578174e-10  4638.646692 0.000000e+00
Load^2    -3.160000e-15 5.000000e-17   -64.950174 0.000000e+00
```

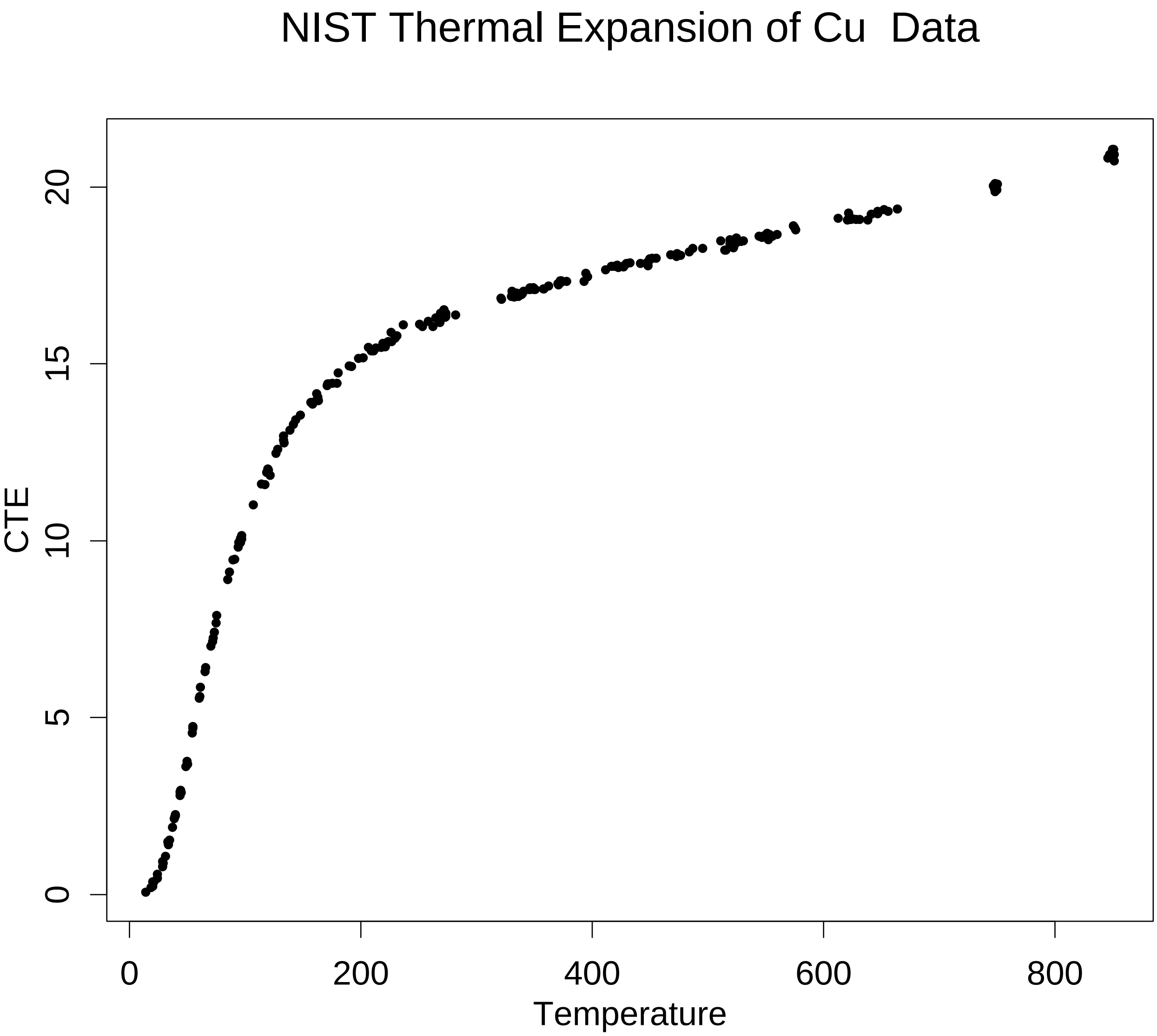

Residuals From Cubic Fit
Residuals
0.0004
0.0002
0.0
-0.0002
-0.0004
5*10^5
10^6
1.5*10^6
2.5*10^6
Load


Residuals From Cubic Fit
Residuals
0.0004
0.0002
0.0
-0.0002
-0.0004
0.5
1.0
1.5
2.0
Predicted Values


Residuals From Cubic Fit
Run Order Plot
Residual
0.0004
0.0
-0.0004
0
10
20
30
40
Observation Number
Lag Plot
Residual Lag 1
0.0004
0.0
-0.0004
-0.0004
0.0
0.0004
Residual
Histogram
Frequency
0
2
4
6
8
10
-0.0004
0.0
0.0004
Residual
Normal Probability Plot
Residual
0.0004
0.0
-0.0004
-2
-1
0
1
2
Quantiles of Standard Normal

Load Cell Data Regression Output
Cubic Model

```
N = 40

Residual Standard Error = 0.0002046495

Multiple R-Square = 0.9999999

F-statistic = 124192184 on 3 and 36 df, p-value = 0

                   coef      std.err       t.stat      p.value
Intercept  5.472497e-04 1.580703e-04     3.462065 1.399291e-03
Load       7.324889e-07 4.240109e-10  1727.523597 0.000000e+00
Load^2    -3.490000e-15 3.100000e-16   -11.313161 2.080600e-13
Load^3     0.000000e+00 0.000000e+00     1.091394 2.823505e-01
```


NIST Thermal Expansion of Cu Data
CTE
0
5
10
15
20
0
200
400
600
800
Temperature

Starting Values for Rational Models

A major advantage of rational models is the ability to compute starting values using a linear least squares fit.

To do this, choose p points from the data set, where p is the number of parameters in the rational model.

Do a linear fit on the 'predictor variables' $x, x^2, \ldots, x^{p_n}, -xy, -x^2y, \ldots, -x^{p_d}y$ and the response variable y, where x and y are the predictor and response variable values selected from the complete data set and p_n and p_d are the degrees of the numerator and denominator of the rational function.

The estimated parameters from the linear fit are your starting values for the nonlinear regression!

Example

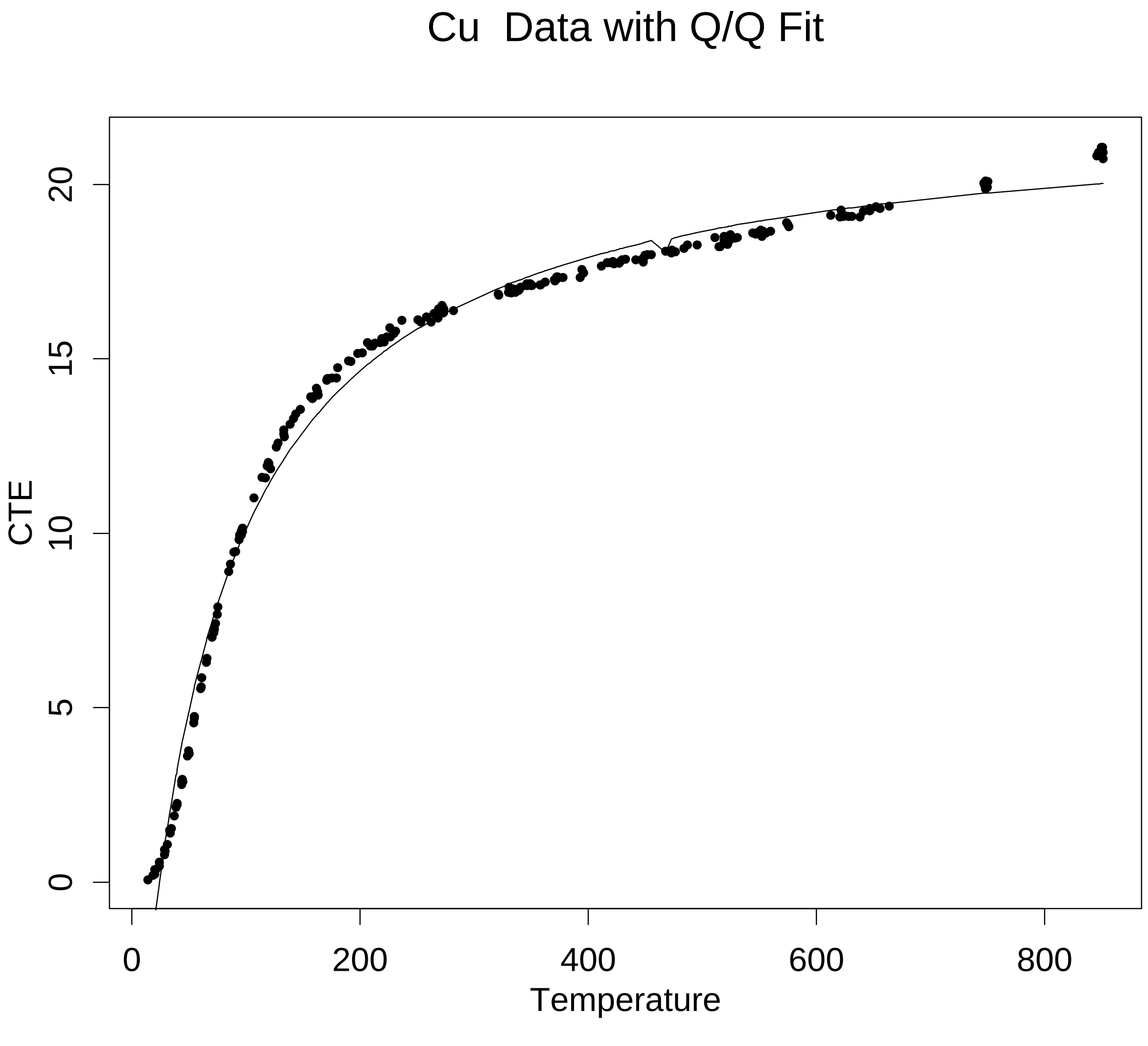
For the Q/Q fit to the Cu data, let x =(18.97, 107.32, 202.14, 495.47, 851.37).

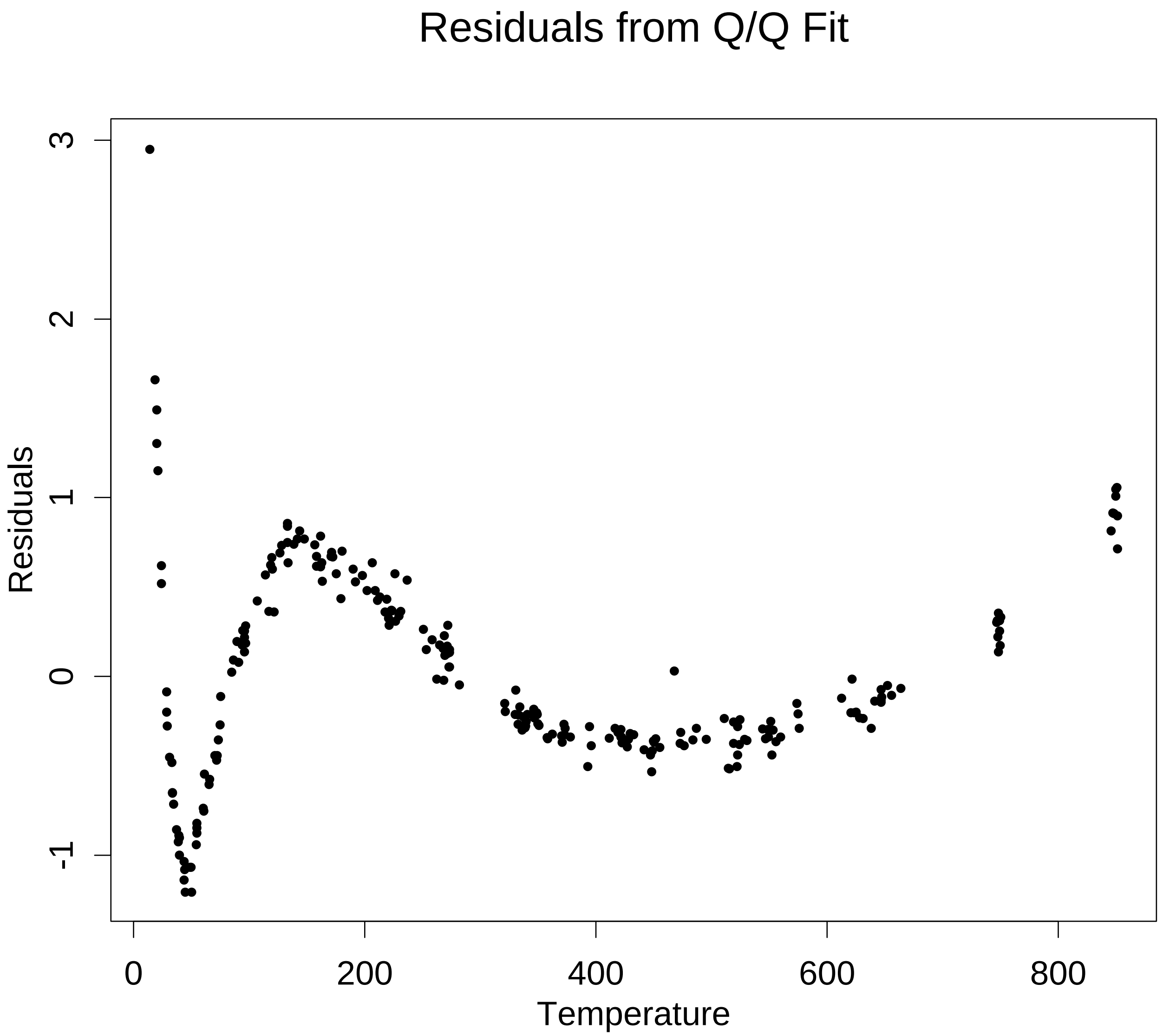
The corresponding values of y are 0.214, 11.023, 15.190, 18.271, and 20.743.

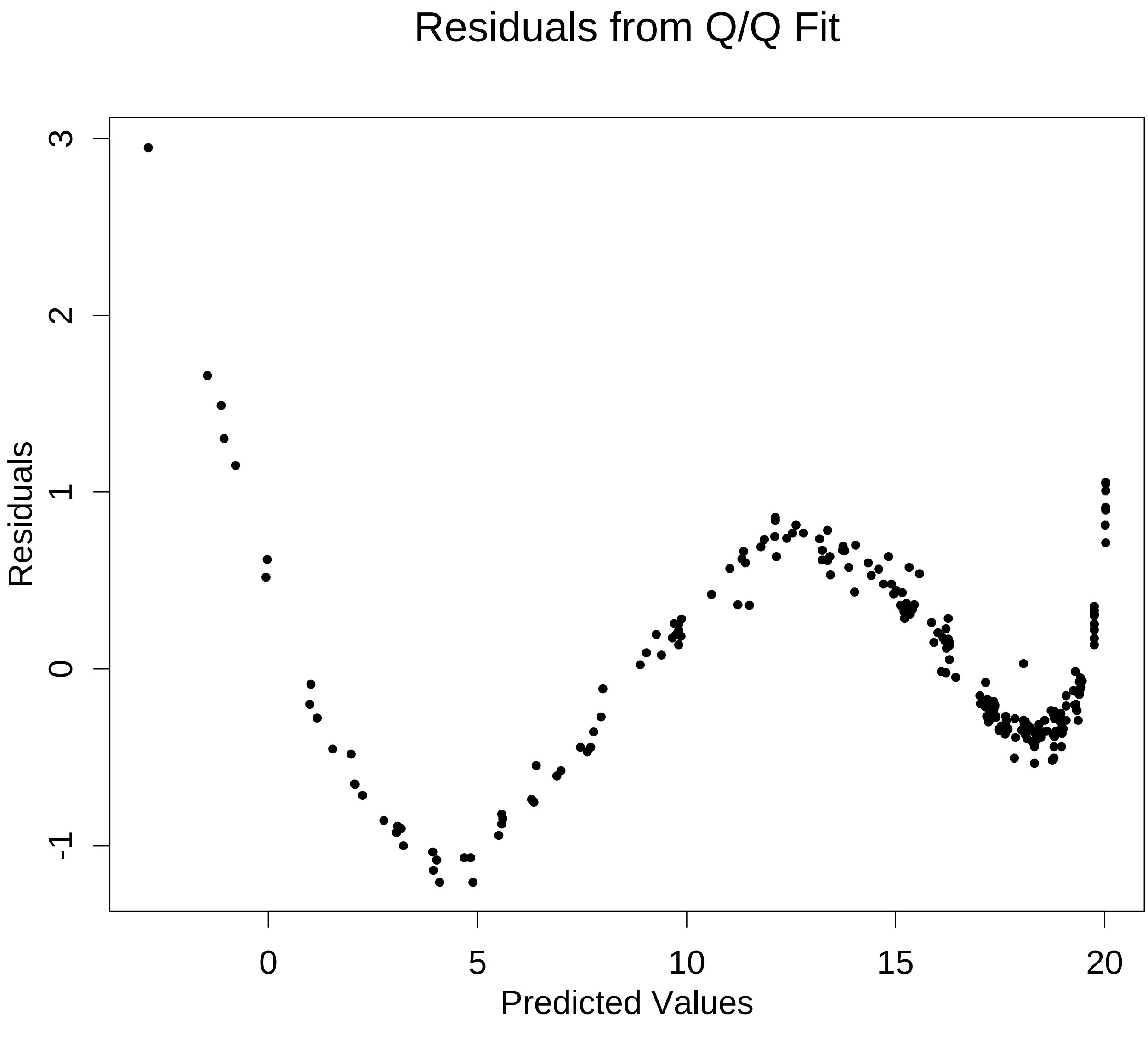
Fitting the function

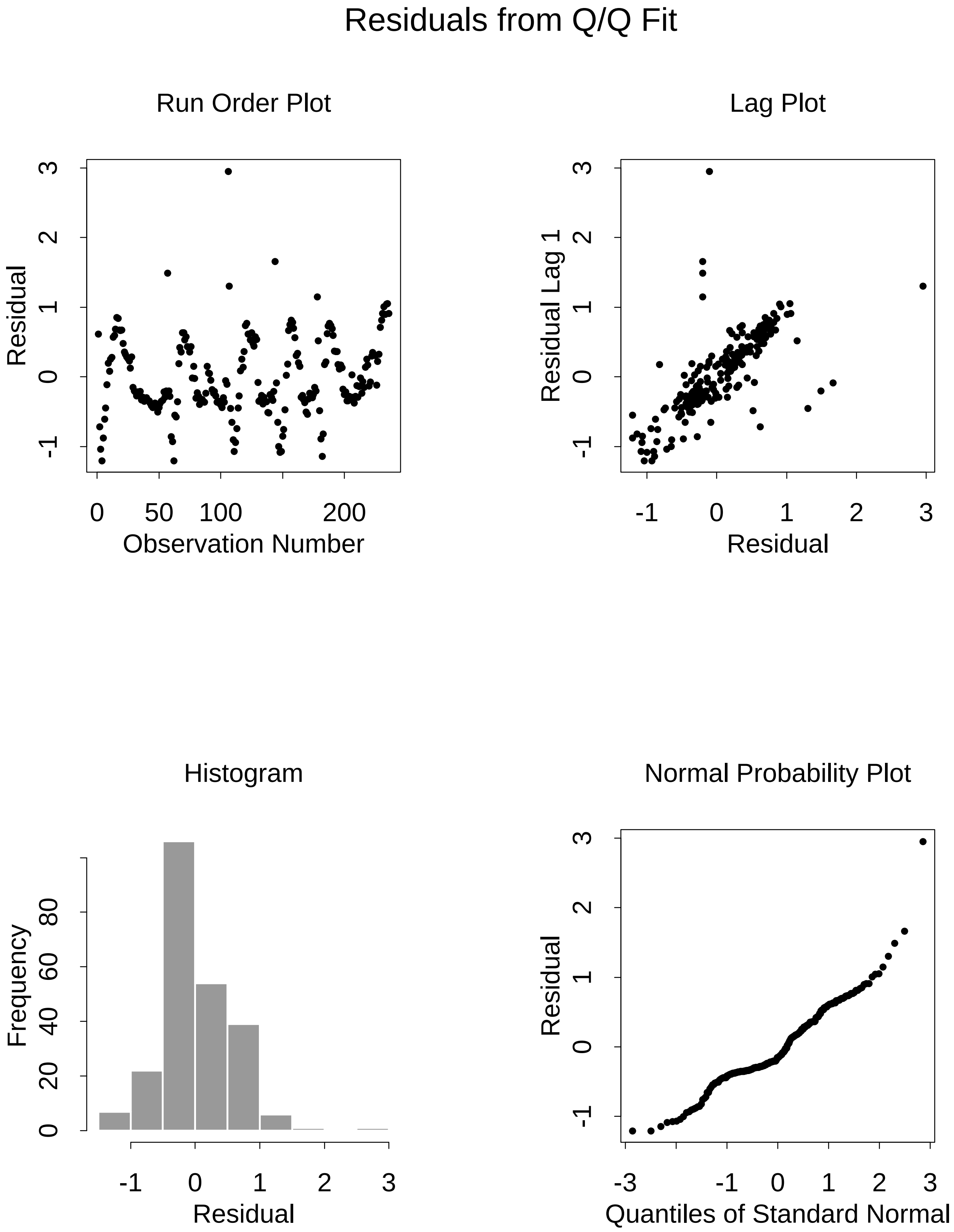
$$y = \beta_1 + \beta_2 x + \beta_3 x^2 - \beta_4 xy - \beta 5 x^2 y$$

to these 5 points yields starting values of -5.11, 0.294, -0.000609, 0.00993, and -2.61e-05 for the Q/Q fit.


Cu Data with Q/Q Fit
CTE
0
5
10
15
20
0
200
400
600
800
Temperature

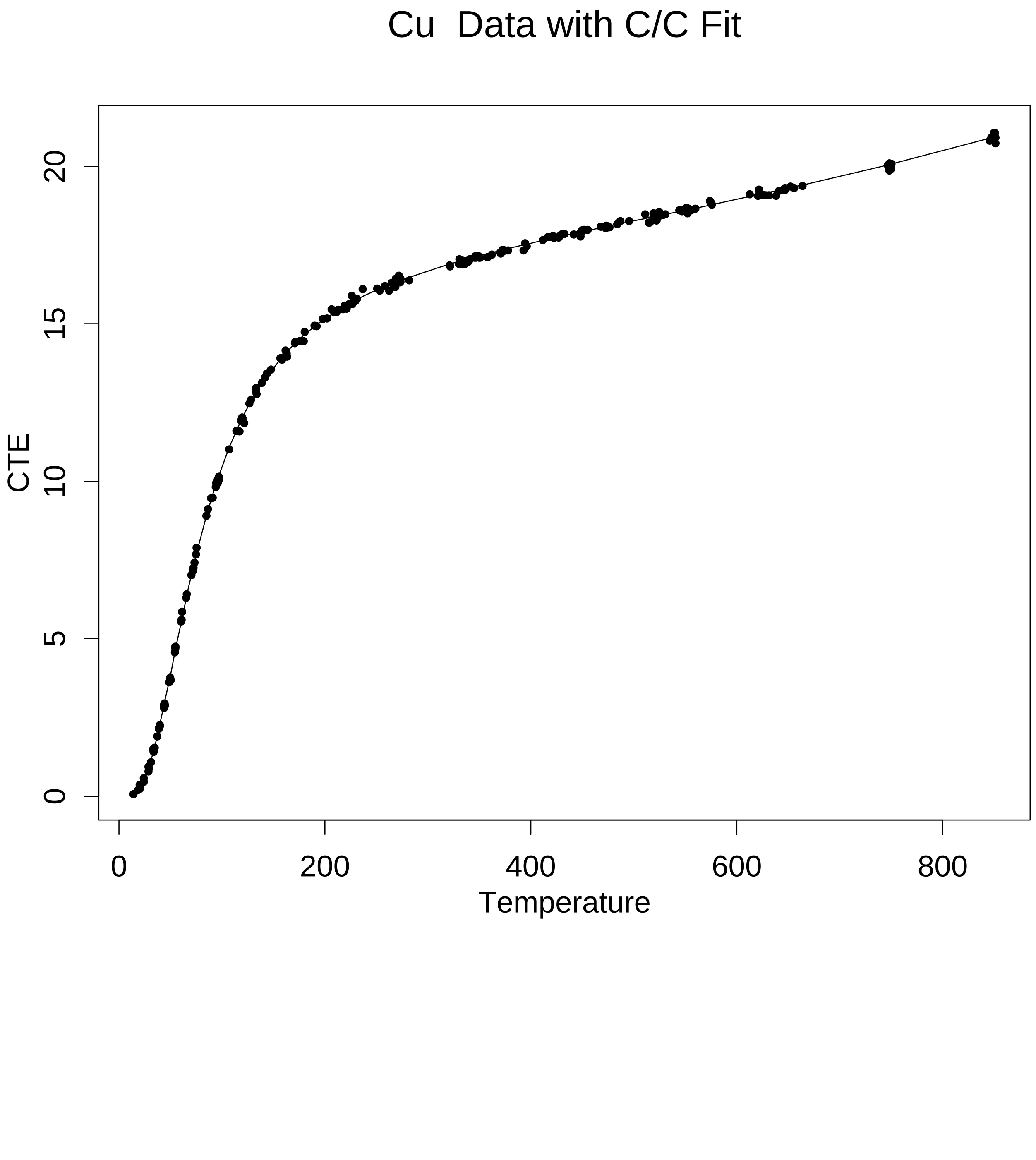

Residuals from Q/Q Fit
Residuals
-1
0
1
2
3
0
200
400
600
800
Temperature

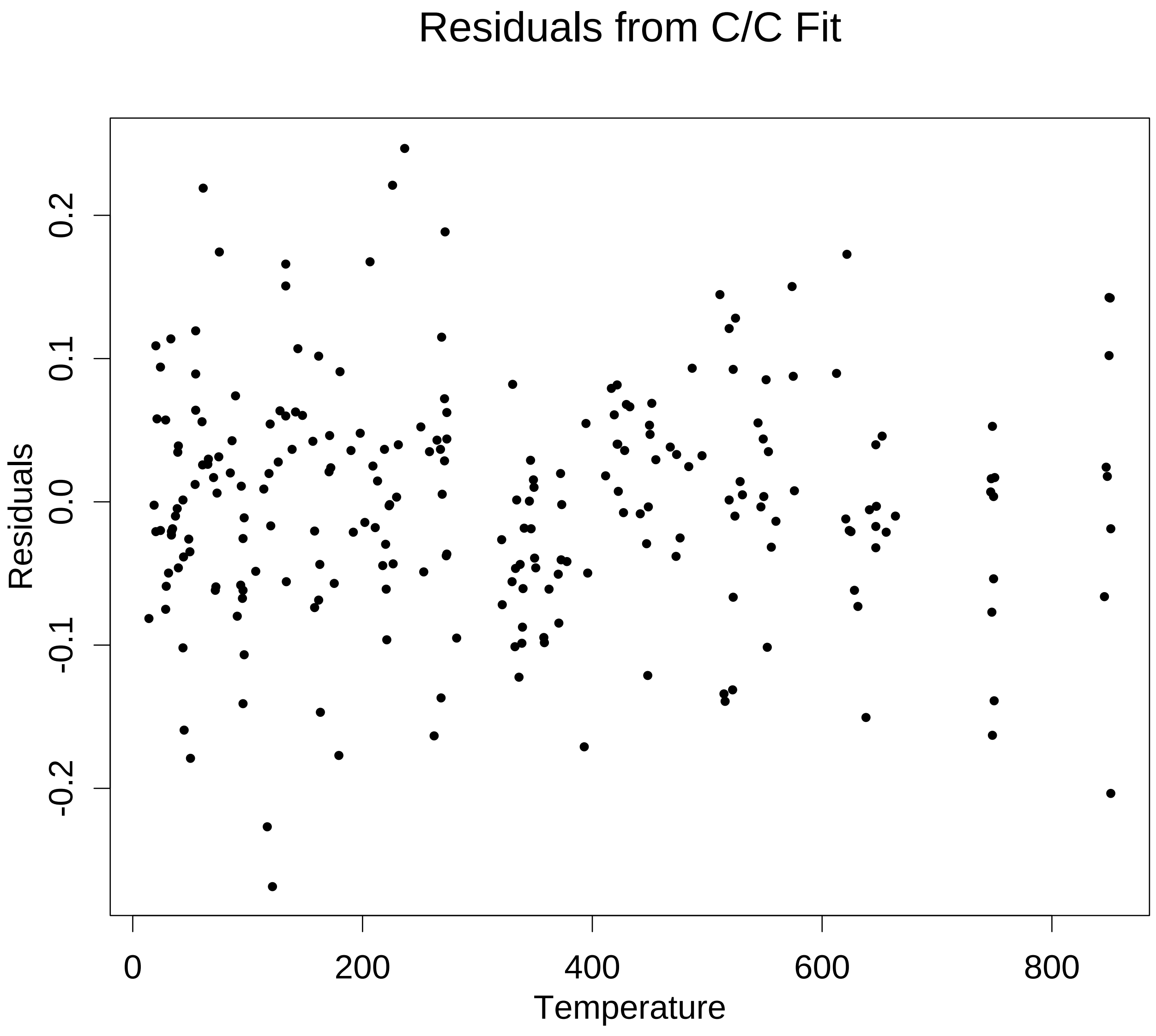

Residuals from Q/Q Fit
Residuals
3
2
1
0
-1
Predicted Values
0
5
10
15
20

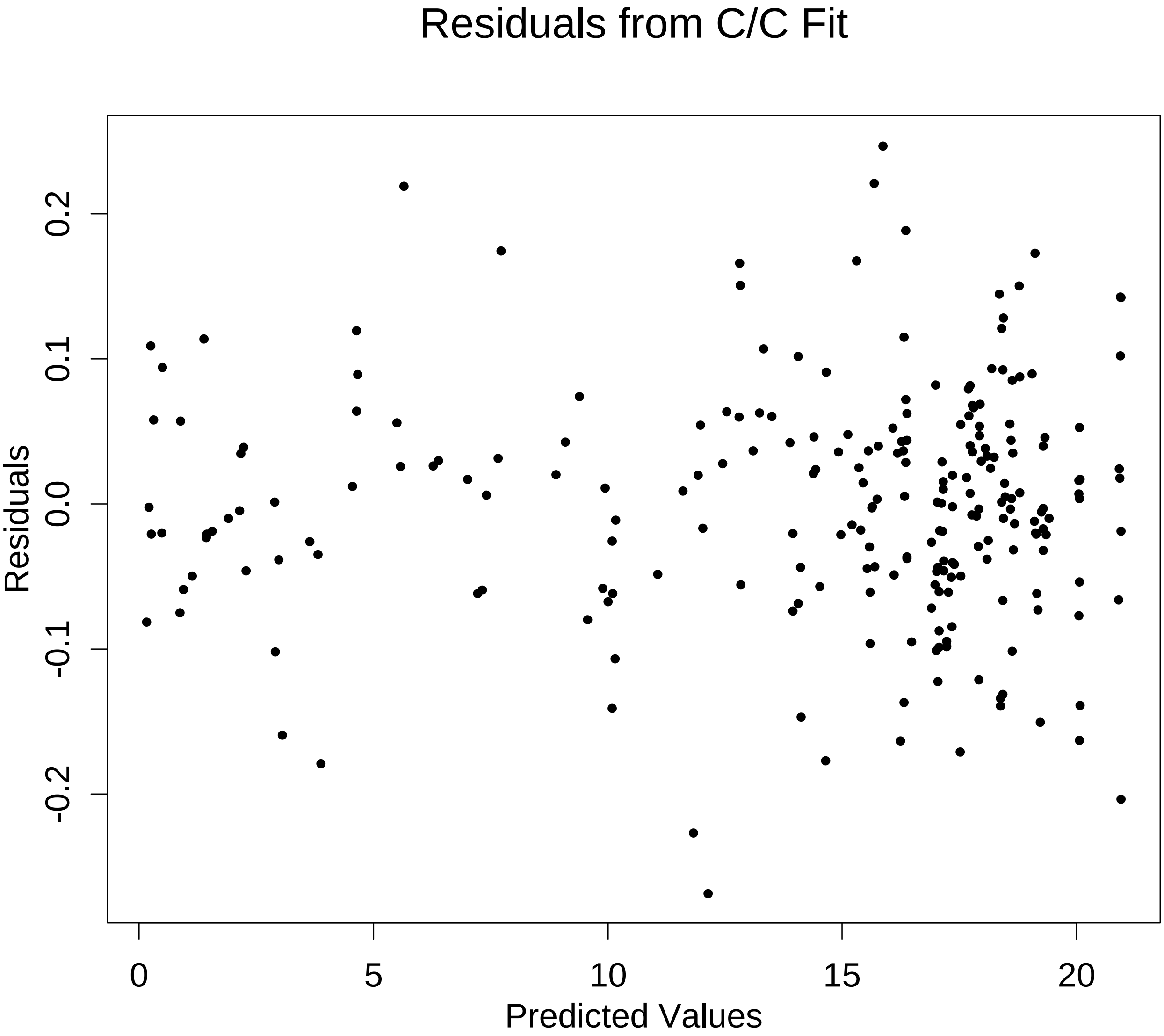

Residuals from Q/Q Fit
Run Order Plot
Residual
-1
0
1
2
3
0
50
100
200
Observation Number
Lag Plot
Residual Lag 1
-1
0
1
2
3
-1
0
1
2
3
Residual
Histogram
Frequency
0
20
40
60
80
-1
0
1
2
3
Residual
Normal Probability Plot
Residual
-1
0
1
2
3
-3
-1
0
1
2
3
Quantiles of Standard Normal

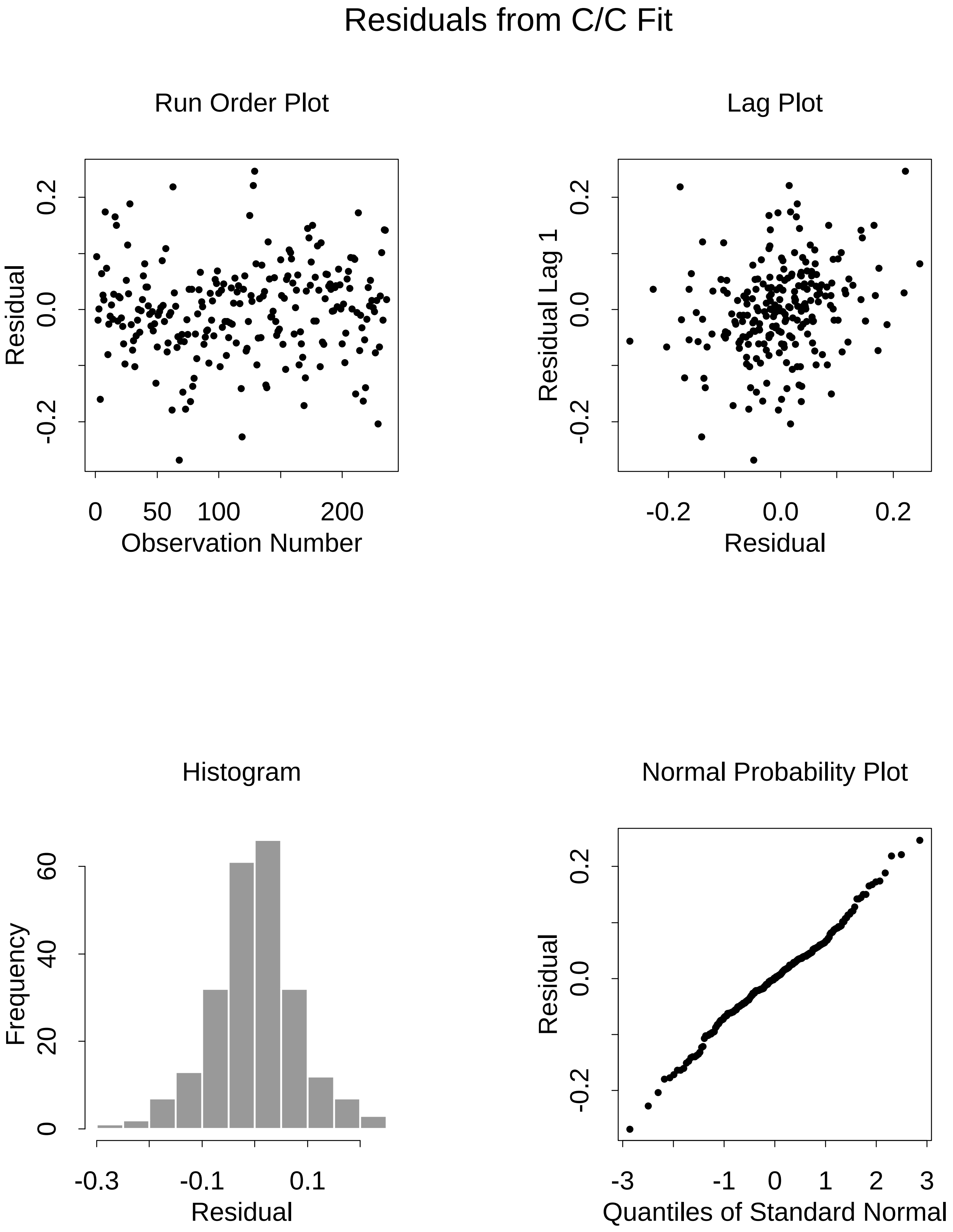
Cu Data Output
Q/Q Model

```
Formula: cte ~ (b1 + b2 * tmp + b3 * tmp^2)/
               (1 + b4 * tmp + b5 * tmp^2)

Parameters:
         Value  Std. Error   t value
b1 -8.21834e+00 3.99196e-01 -20.5872
b2  3.52581e-01 1.11990e-02  31.4833
b3 -7.16809e-04 2.26561e-05 -31.6387
b4  1.29286e-02 5.36350e-04  24.1048
b5 -3.22441e-05 1.15327e-06 -27.9588

Residual standard error: 0.562522 on 231 degrees of freedom

Correlation of Parameter Estimates:
      b1     b2     b3     b4
b2 -0.954
b3  0.905 -0.956
b4 -0.927  0.993 -0.933
b5  0.897 -0.961  0.995 -0.949
```


Cu Data with C/C Fit
CTE
0
5
10
15
20
Temperature
0
200
400
600
800

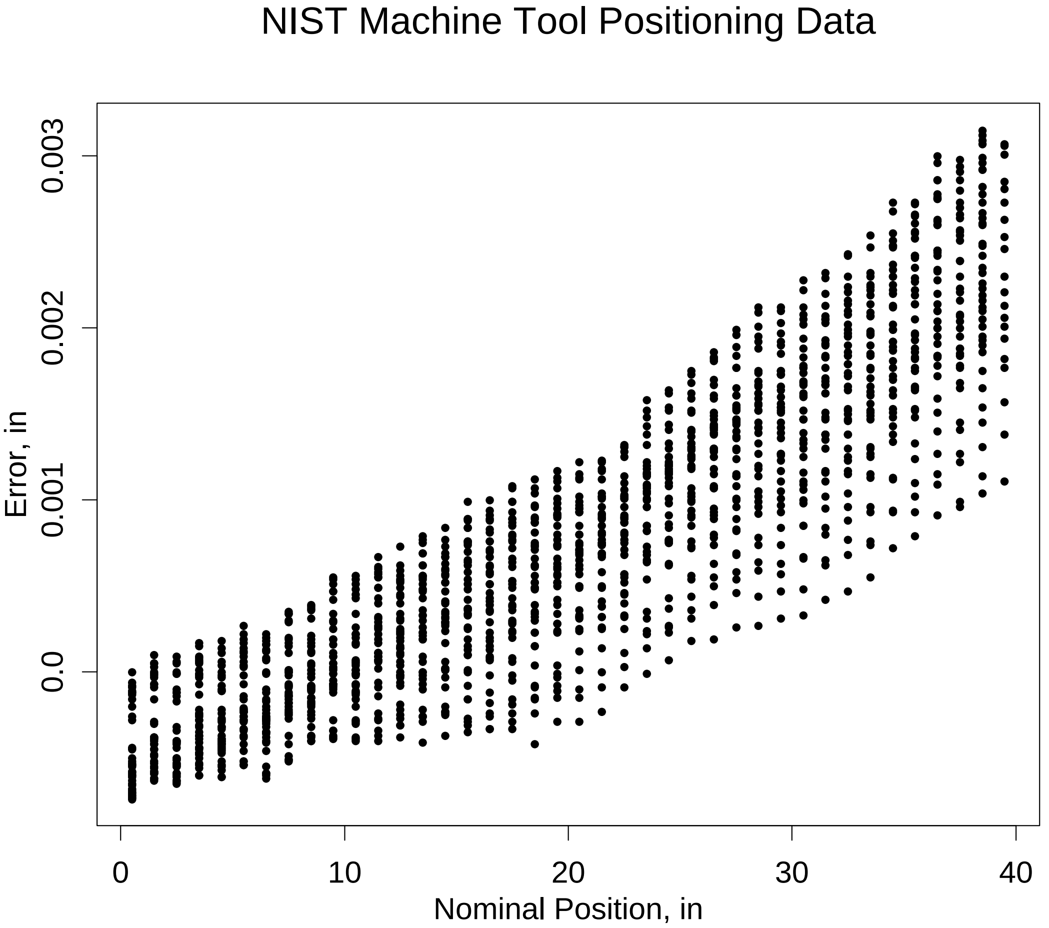

Residuals from C/C Fit
Residuals
0.2
0.1
0.0
-0.1
-0.2
Temperature
0
200
400
600
800

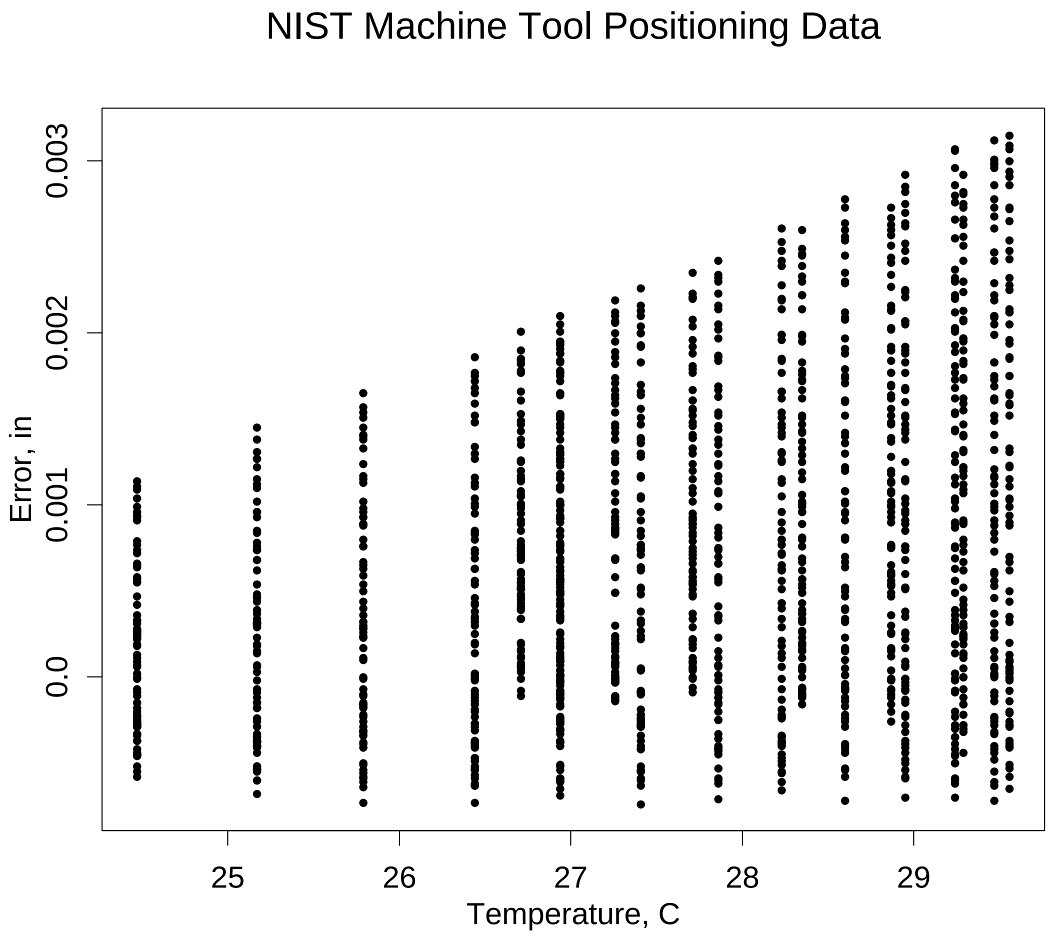

Residuals from C/C Fit
Residuals
0.2
0.1
0.0
-0.1
-0.2
0
5
10
15
20
Predicted Values

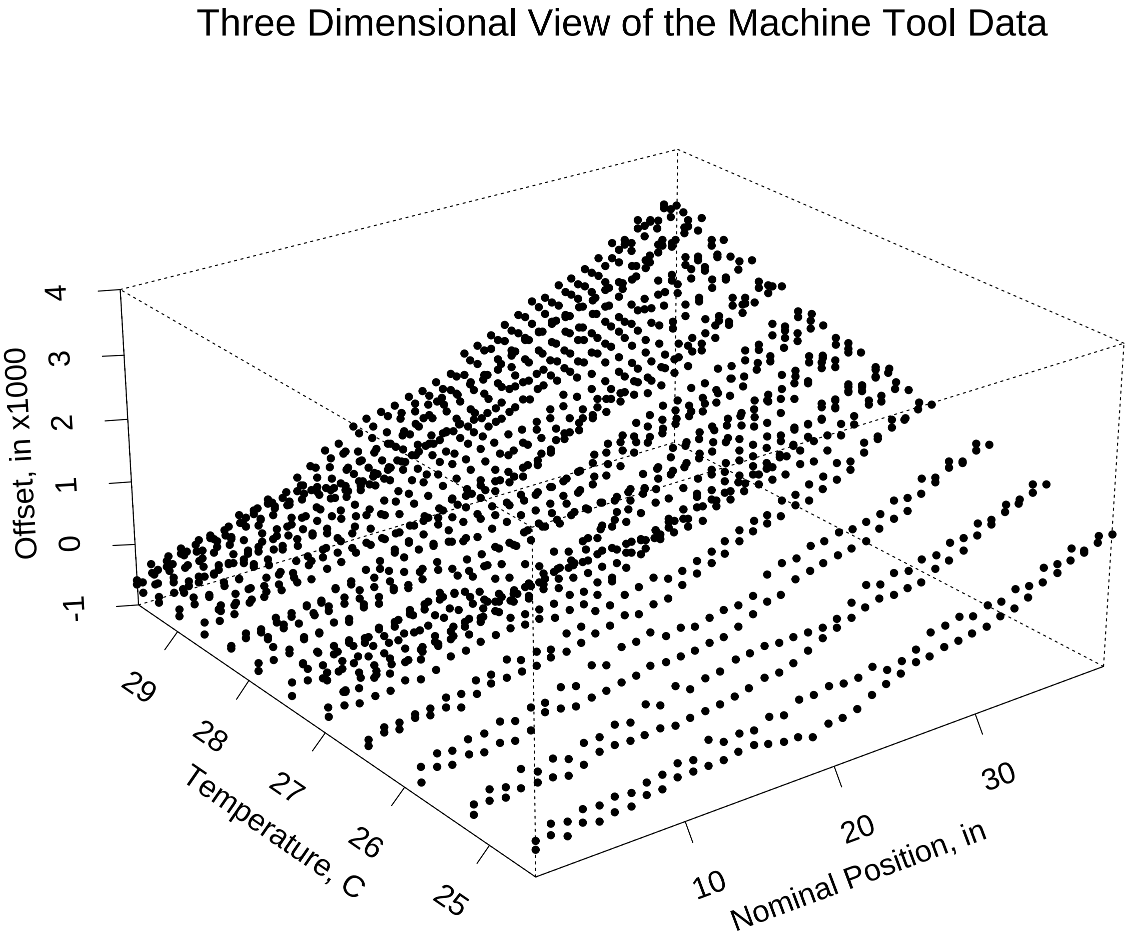

Residuals from C/C Fit
Run Order Plot
Residual
-0.2
0.0
0.2
0
50
100
200
Observation Number
Lag Plot
Residual Lag 1
-0.2
0.0
0.2
-0.2
0.0
0.2
Residual
Histogram
Frequency
0
20
40
60
-0.3
-0.1
0.1
Residual
Normal Probability Plot
Residual
-0.2
0.0
0.2
-3
-1
0
1
2
3
Quantiles of Standard Normal

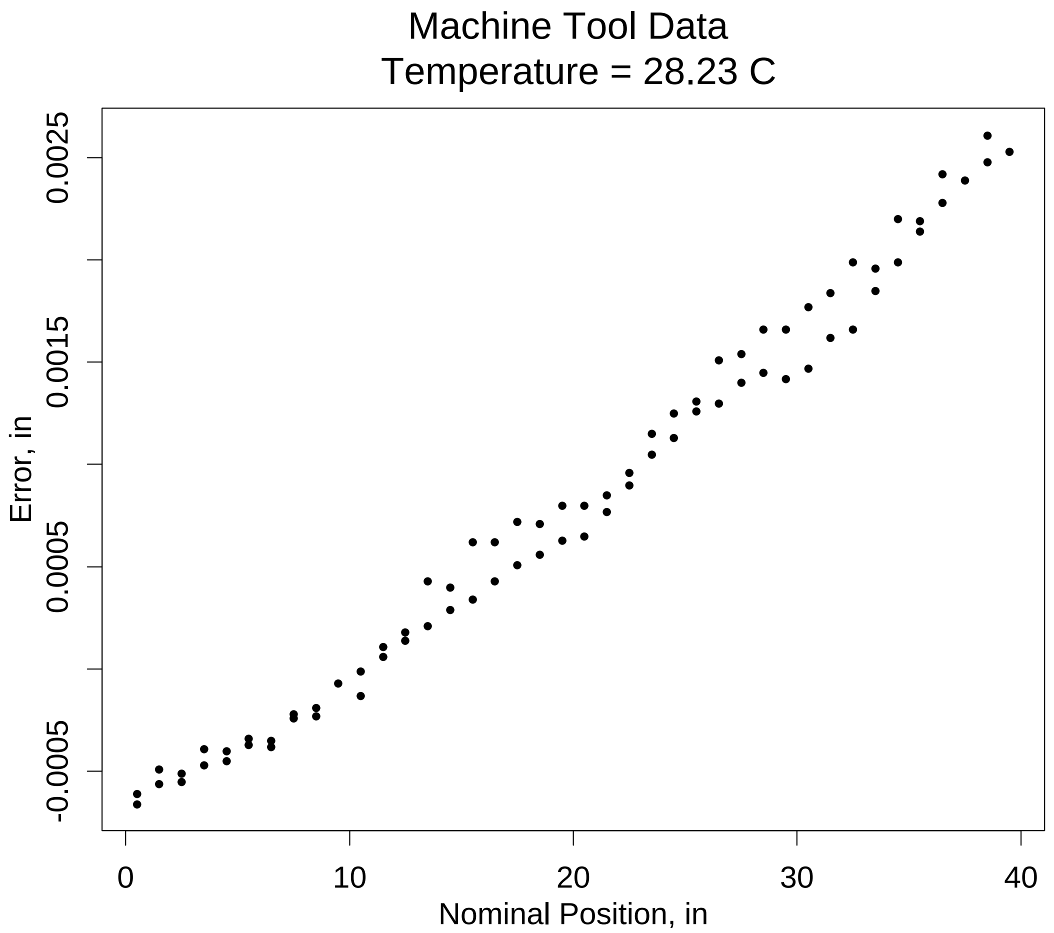
Cu Data Output
C/C Model

```
Formula: cte ~ (b1 + b2 * tmp + b3 * tmp^2 + b4 * tmp^3)/
               (1 + b5 * tmp + b6 * tmp^2 + b7 * tmp^3)

Parameters:
         Value  Std. Error   t value
b1  1.07766e+00 1.70702e-01   6.31312
b2 -1.22695e-01 1.20004e-02 -10.22430
b3  4.08642e-03 2.25085e-04  18.15500
b4 -1.42632e-06 2.75781e-07  -5.17193
b5 -5.76099e-03 2.47130e-04 -23.31160
b6  2.40539e-04 1.04494e-05  23.01930
b7 -1.23147e-07 1.30274e-08  -9.45295

Residual standard error: 0.0818039 on 229 degrees of freedom

Correlation of Parameter Estimates:
       b1       b2       b3       b4       b5       b6
b2 -0.97200
b3  0.90500 -0.97700
b4 -0.70300  0.79600 -0.87000
b5 -0.26300  0.09530  0.10800 -0.40200
b6  0.92500 -0.98600  0.99300 -0.80900  0.00861
b7 -0.75900  0.84700 -0.91000  0.99500 -0.33200 -0.86000
```

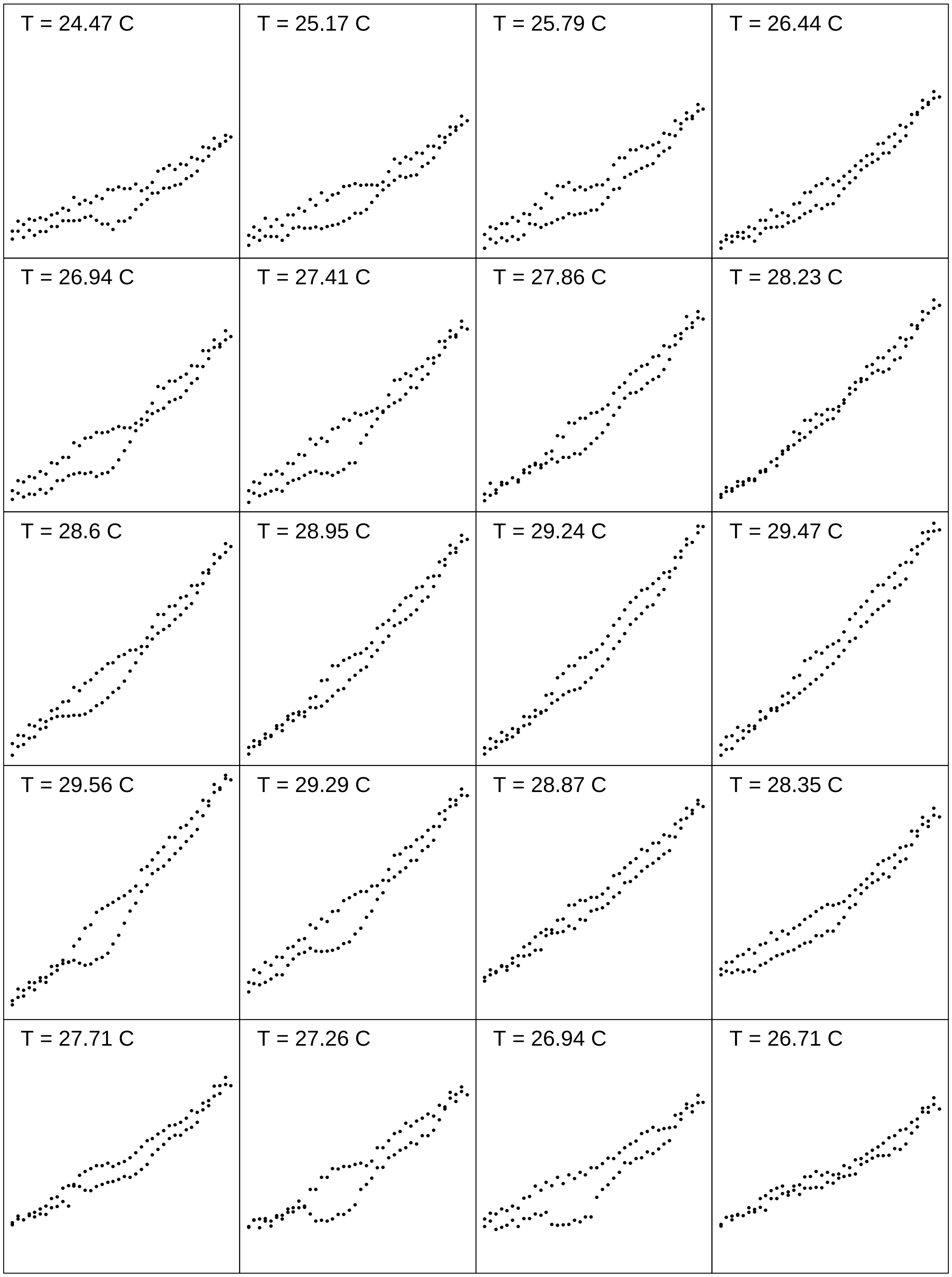

NIST Machine Tool Positioning Data
Error, in
0.0
0.001
0.002
0.003
0
10
20
30
40
Nominal Position, in

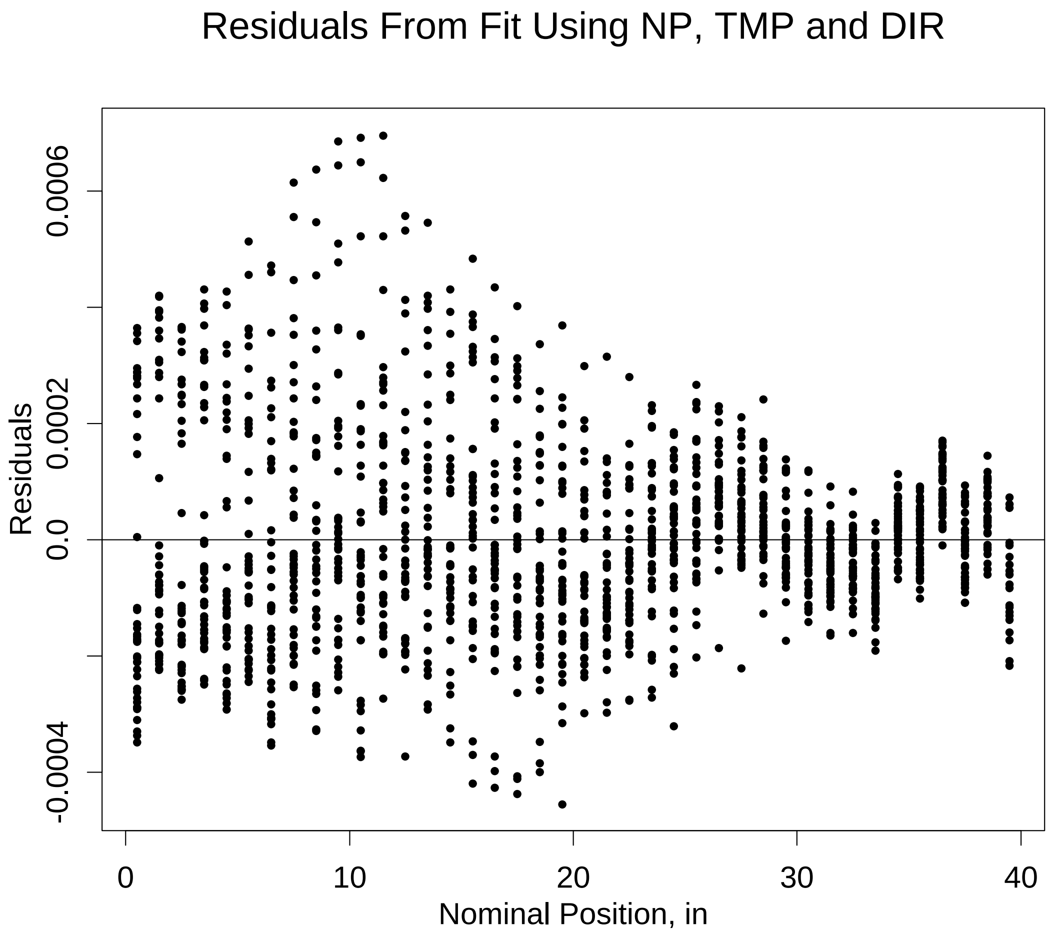

NIST Machine Tool Positioning Data
Error, in
0.0
0.001
0.002
0.003
Temperature, C
25
26
27
28
29

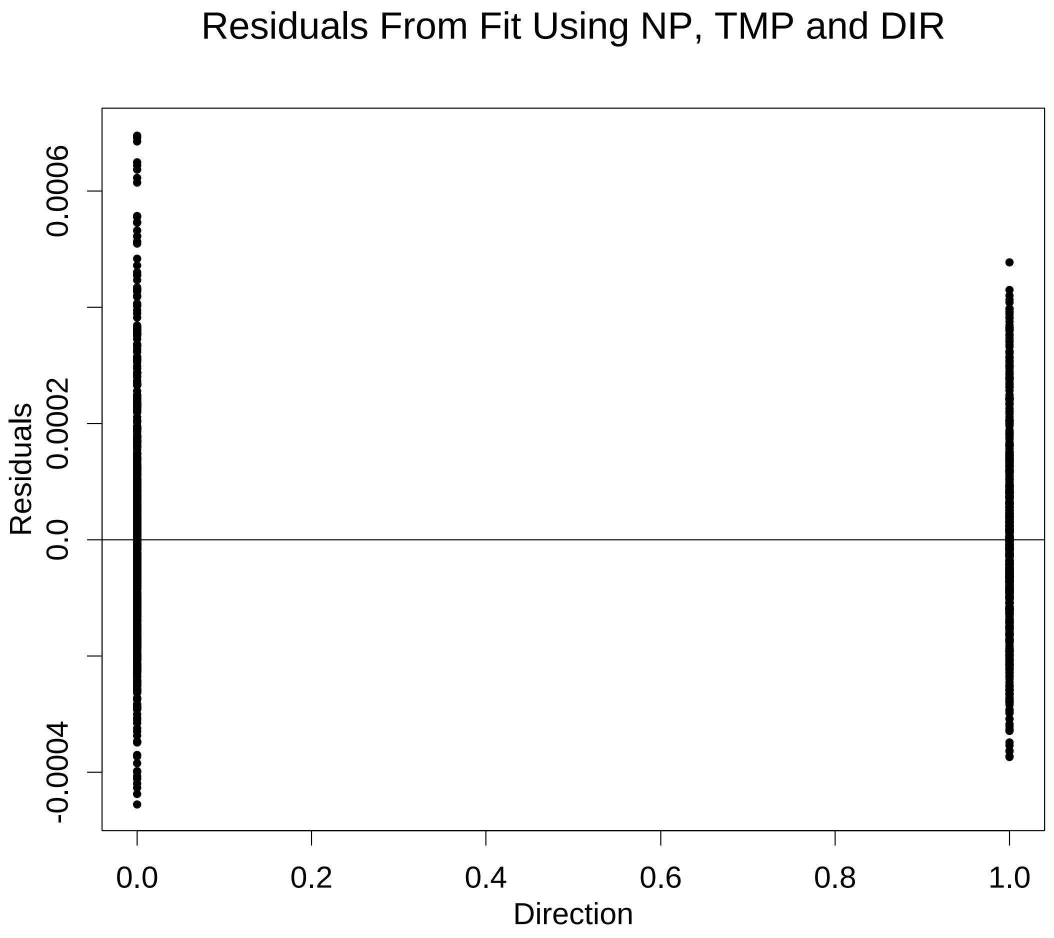

Three Dimensional View of the Machine Tool Data
Offset, in x1000
-1
0
1
2
3
4
29
28
27
26
25
Temperature, C
10
20
30
Nominal Position, in

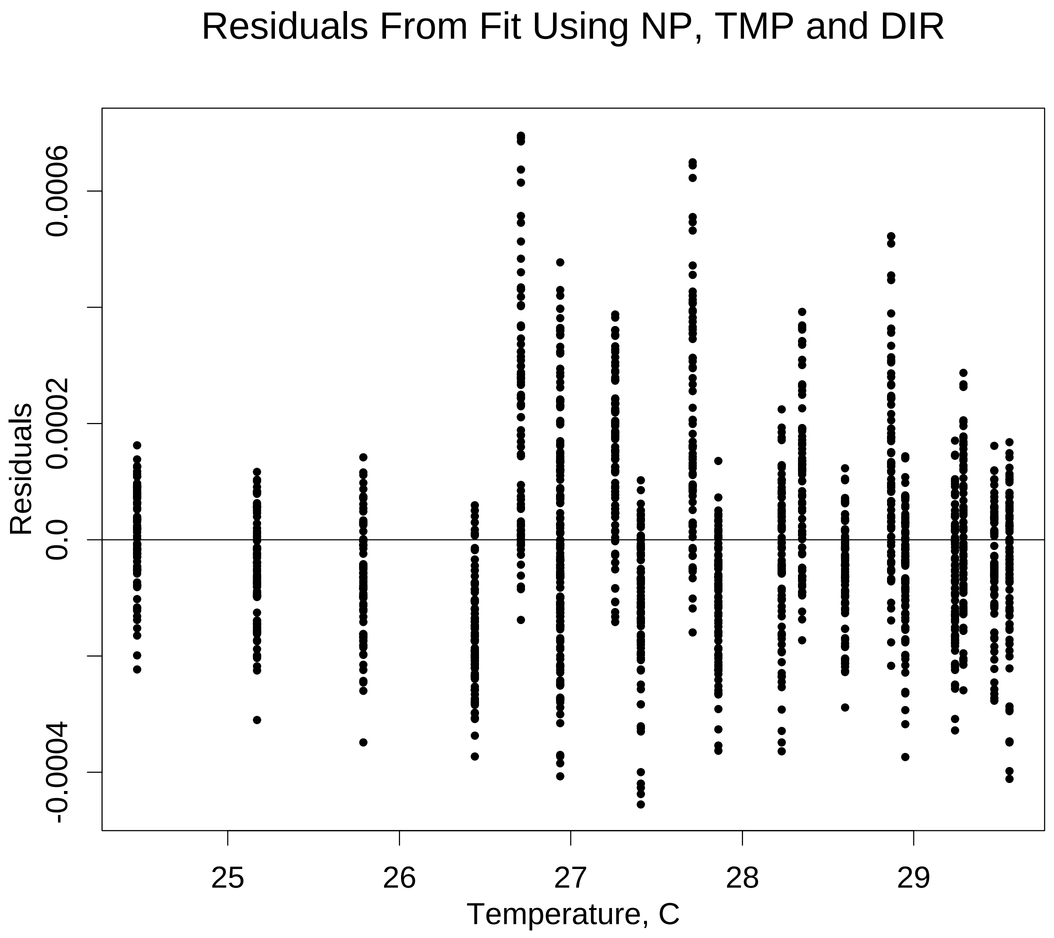

Machine Tool Data
Temperature = 28.23 C
Error, in
-0.0005
0.0005
0.0015
0.0025
0
10
20
30
40
Nominal Position, in

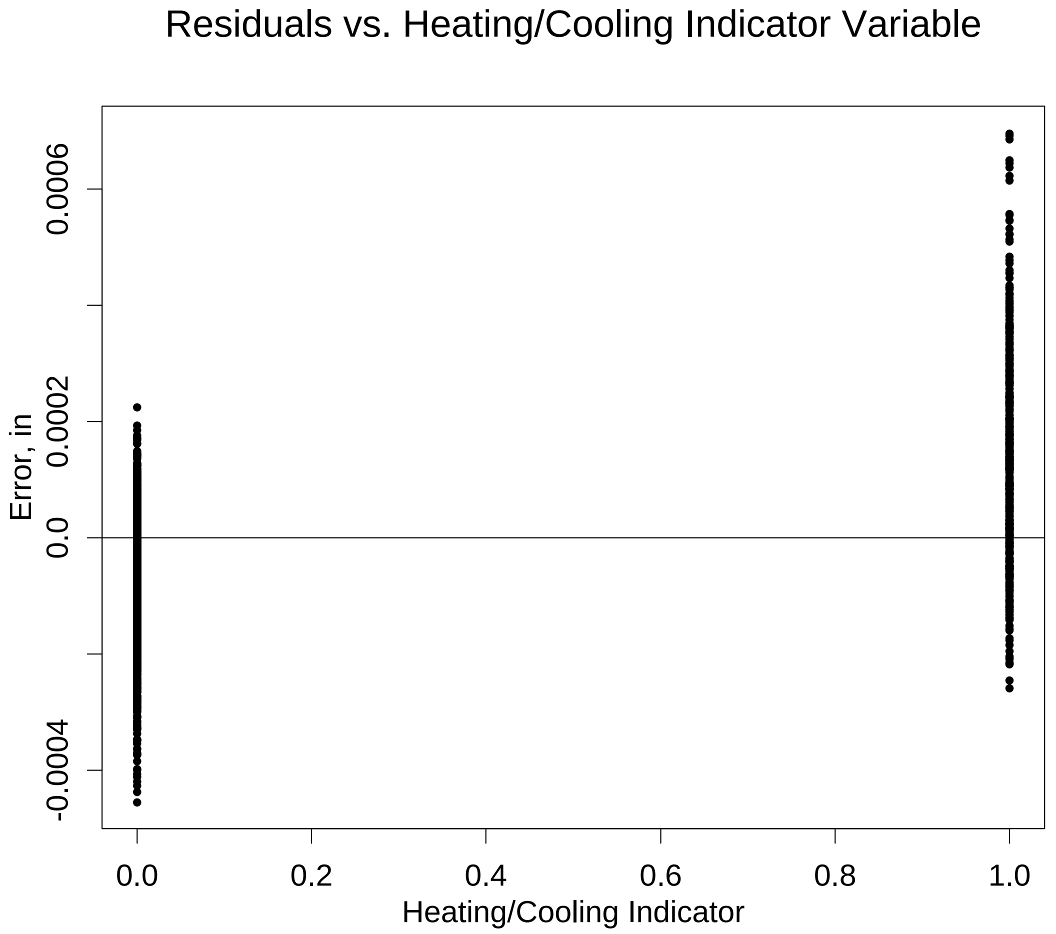
Cross-Sectional Plots of Machine Tool Data
Fixed Scale for X & Y Axes

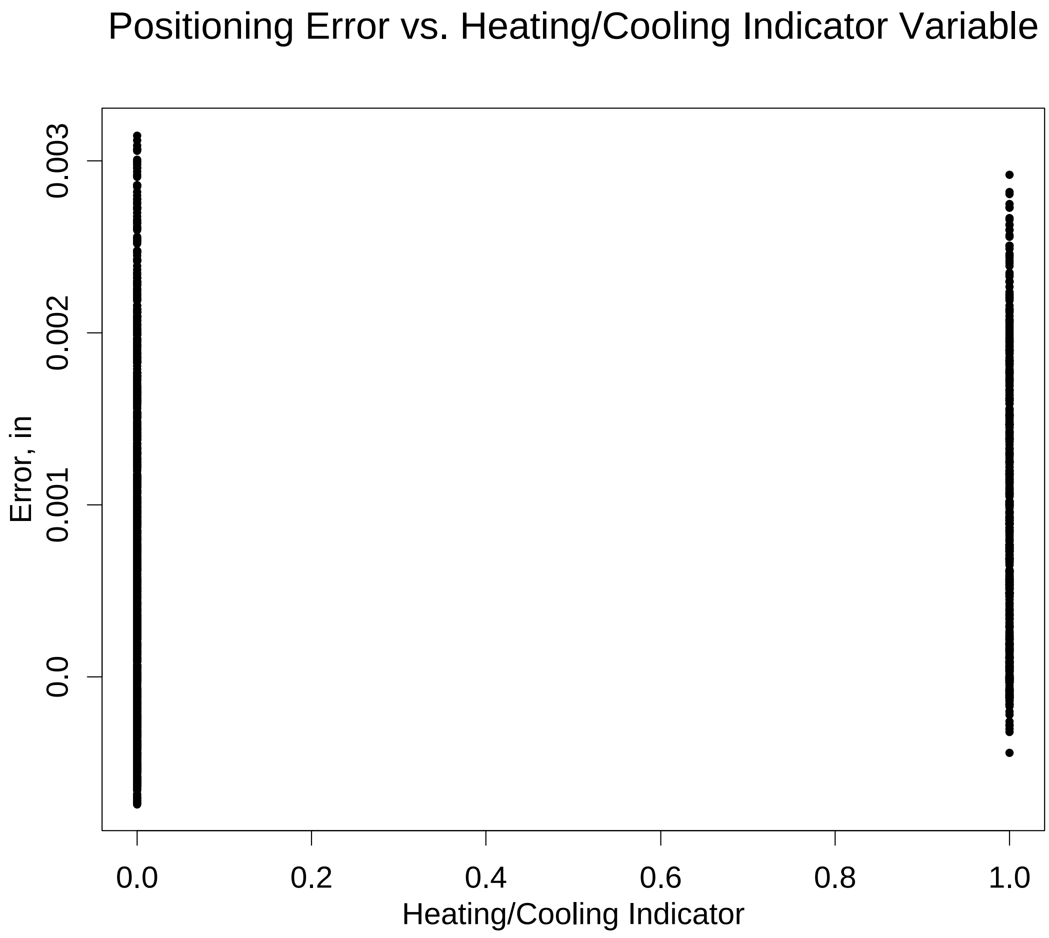


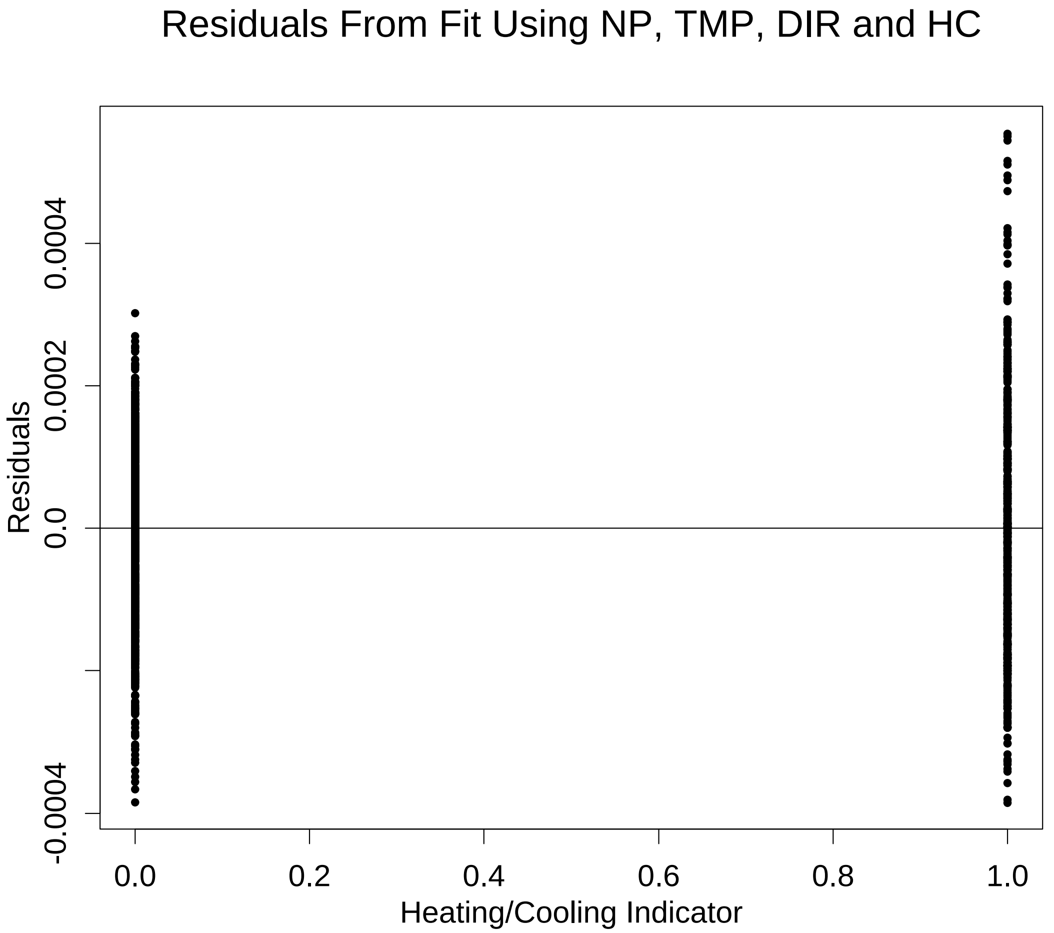

Residuals From Fit Using NP, TMP and DIR
Residuals
0.0006
0.0002
0.0
-0.0004
0
10
20
30
40
Nominal Position, in

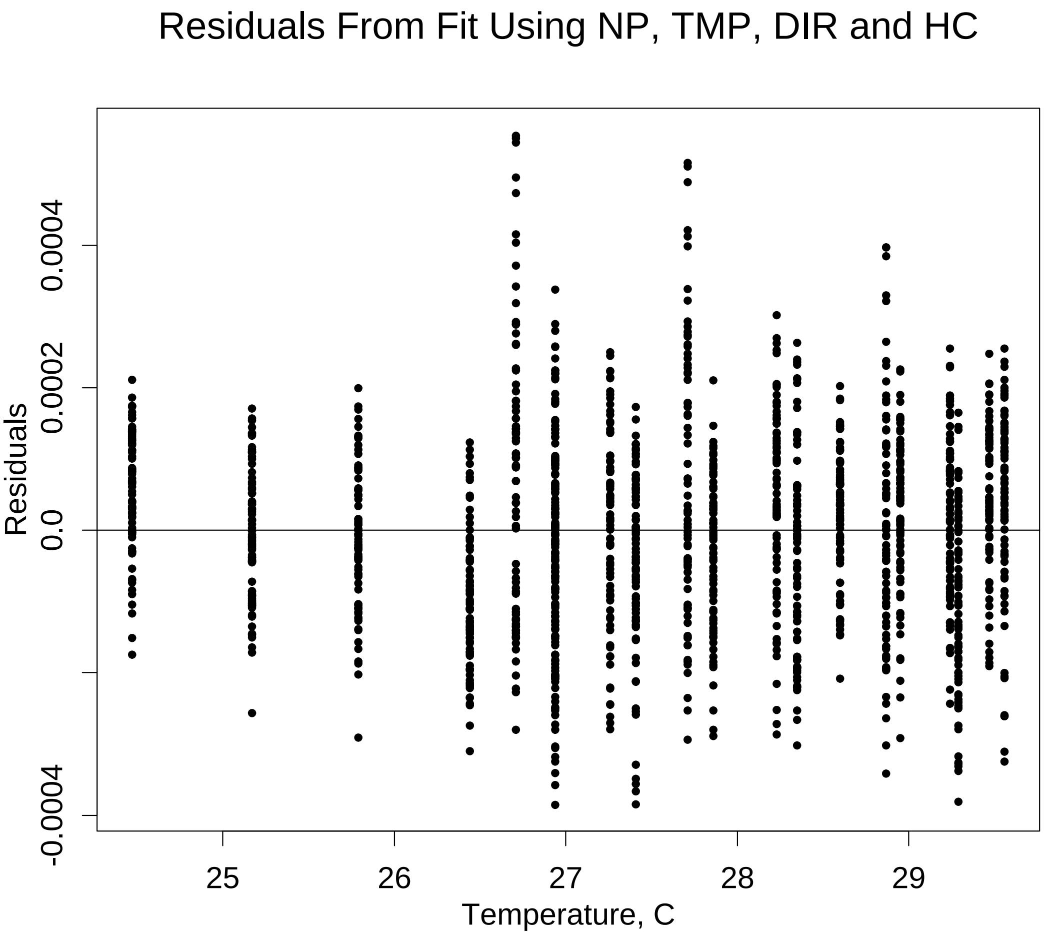

Residuals From Fit Using NP, TMP and DIR
Residuals
0.0006
0.0002
0.0
-0.0004
0.0
0.2
0.4
0.6
0.8
1.0
Direction

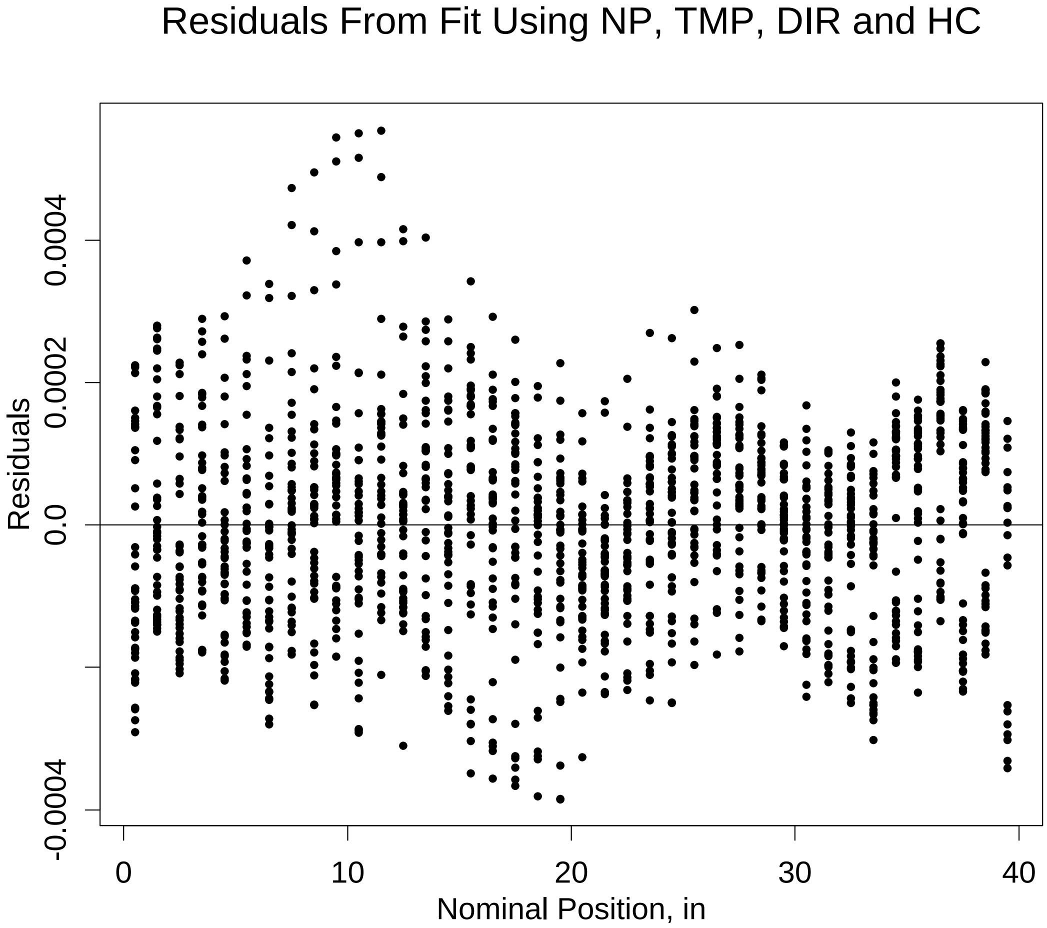

Residuals From Fit Using NP, TMP and DIR
Residuals
0.0006
0.0002
0.0
-0.0004
25
26
27
28
29
Temperature, C

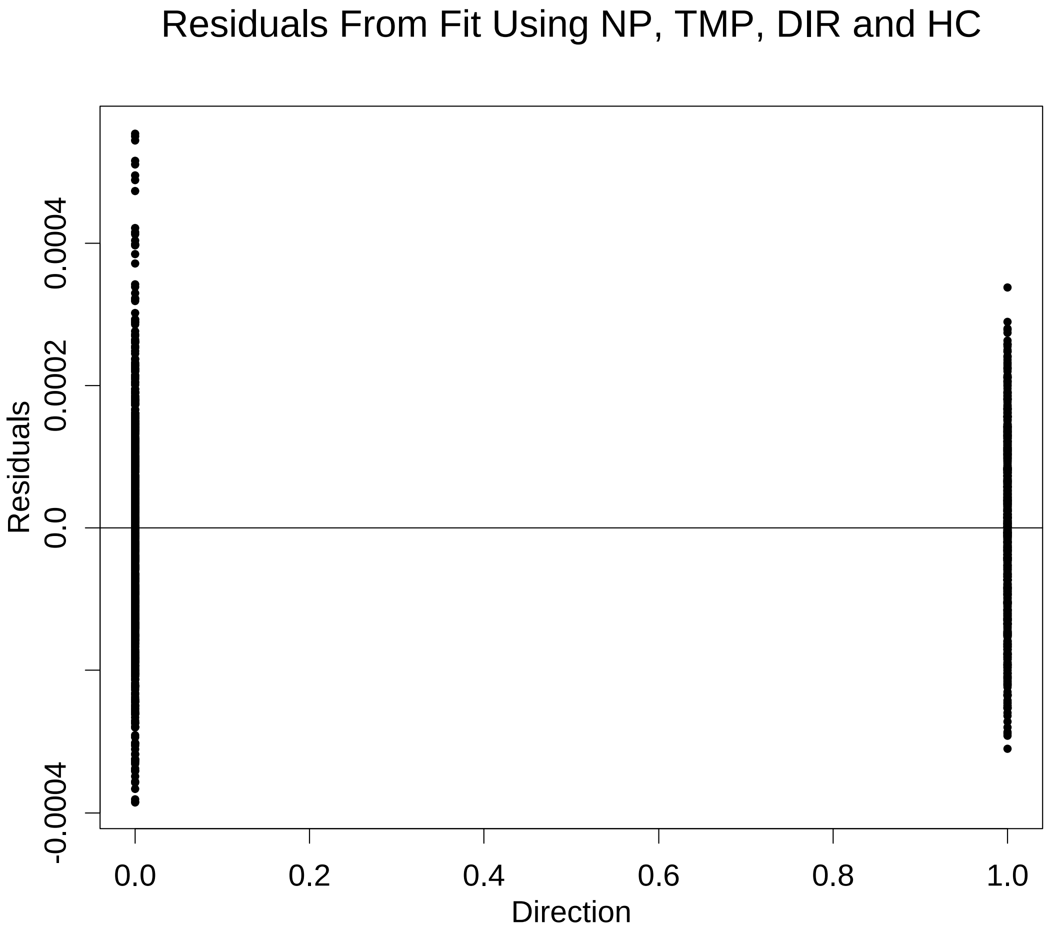

Residuals vs. Heating/Cooling Indicator Variable
Error, in
0.0006
0.0002
0.0
-0.0004
0.0
0.2
0.4
0.6
0.8
1.0
Heating/Cooling Indicator

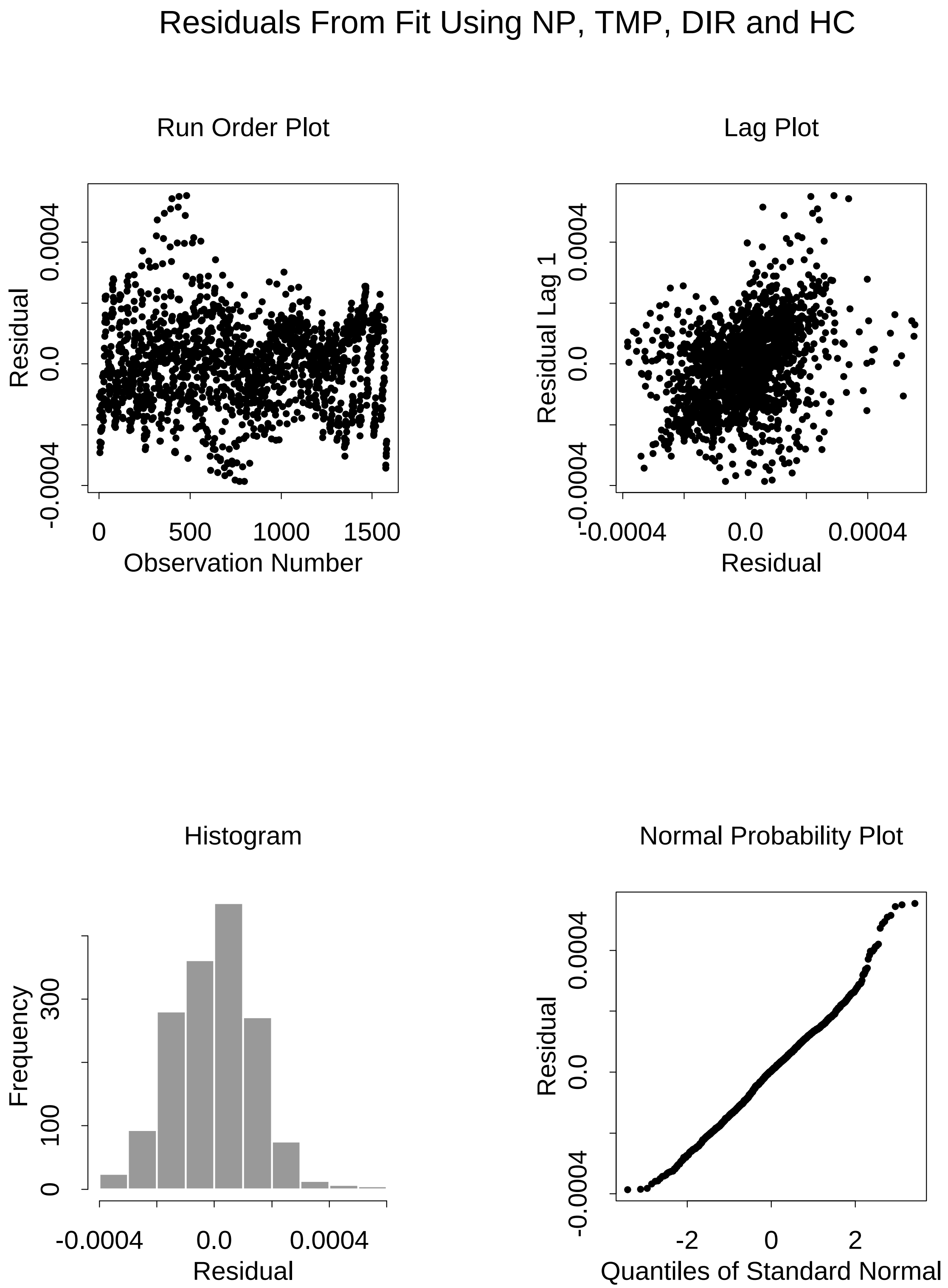

Positioning Error vs. Heating/Cooling Indicator Variable
Error, in
0.0
0.001
0.002
0.003
0.0
0.2
0.4
0.6
0.8
1.0
Heating/Cooling Indicator


Residuals From Fit Using NP, TMP, DIR and HC
Residuals
0.0004
0.0002
0.0
-0.0004
0.0
0.2
0.4
0.6
0.8
1.0
Heating/Cooling Indicator


Residuals From Fit Using NP, TMP, DIR and HC
Residuals
0.0004
0.0002
0.0
-0.0004
25
26
27
28
29
Temperature, C


Residuals From Fit Using NP, TMP, DIR and HC
Residuals
0.0004
0.0002
0.0
-0.0004
0
10
20
30
40
Nominal Position, in


Residuals From Fit Using NP, TMP, DIR and HC
Residuals
0.0004
0.0002
0.0
-0.0004
0.0
0.2
0.4
0.6
0.8
1.0
Direction

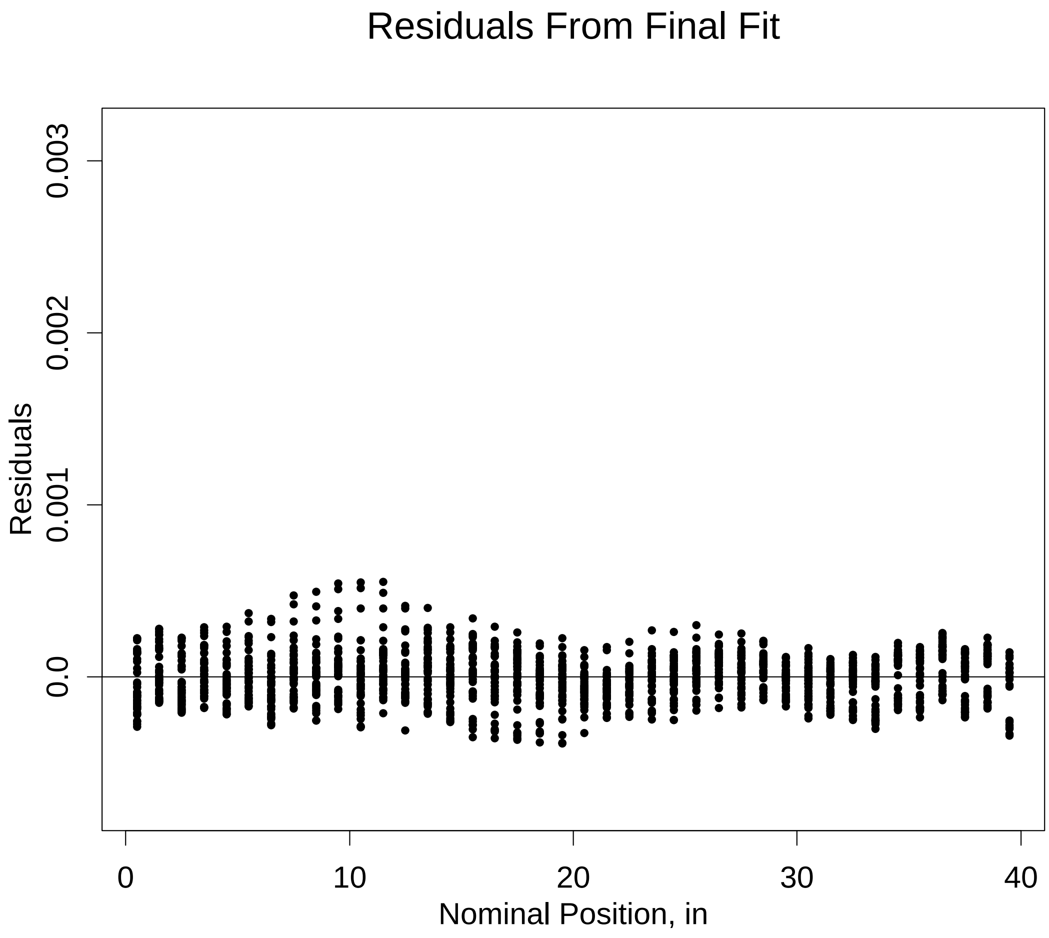

Residuals From Fit Using NP, TMP, DIR and HC
Run Order Plot
Residual
-0.0004
0.0
0.0004
0
500
1000
1500
Observation Number
Lag Plot
Residual Lag 1
-0.0004
0.0
0.0004
-0.0004
0.0
0.0004
Residual
Histogram
Frequency
0
100
300
-0.0004
0.0
0.0004
Residual
Normal Probability Plot
Residual
-0.0004
0.0
0.0004
-2
0
2
Quantiles of Standard Normal

Machine Tool Regression Output

```
N = 1580

Residual Standard Error = 0.0001399666

Multiple R-Square = 0.9760462

F-statistic = 6393.203 on 10 and 1569 df, p-value = 0

                   coef      std.err       t.stat      p.value
Intercept  3.668607e-04 2.165631e-04   1.69401280 9.046132e-02
NP        -2.884838e-04 2.410003e-05 -11.97026619 0.000000e+00
NP^2       1.371996e-06 5.904339e-07   2.32370755 2.026838e-02
TMP       -3.122734e-05 7.812888e-06  -3.99690109 6.715456e-05
DIR        2.618763e-04 1.387625e-04   1.88722618 5.931436e-02
NP*DIR     3.419570e-05 2.475432e-06  13.81403090 0.000000e+00
NP*TMP     1.105673e-05 8.689877e-07  12.72369131 0.000000e+00
TMP*DIR   -8.949689e-06 4.957547e-06  -1.80526564 7.122475e-02
TMP*NP^2  -1.229523e-09 2.129087e-08  -0.05774884 9.539560e-01
DIR*NP^2  -8.437671e-07 6.072783e-08 -13.89424058 0.000000e+00
HC         2.069802e-04 7.427440e-06  27.86695871 0.000000e+00
```

Machine Tool Regression Output

```
N = 1580

Residual Standard Error = 0.0001399222

Multiple R-Square = 0.9760461

F-statistic = 7108.071 on 9 and 1570 df, p-value = 0

                   coef      std.err     t.stat      p.value
Intercept  3.581151e-04 1.547573e-04   2.314043 2.079390e-02
NP        -2.871397e-04 6.248114e-06 -45.956211 0.000000e+00
NP^2       1.337984e-06 4.154677e-08  32.204276 0.000000e+00
TMP       -3.091119e-05 5.572357e-06  -5.547239 3.400336e-08
DIR        2.616527e-04 1.386645e-04   1.886948 5.935163e-02
NP*DIR     3.419570e-05 2.474646e-06  13.818418 0.000000e+00
NP*TMP     1.100814e-05 2.171776e-07  50.687284 0.000000e+00
TMP*DIR   -8.941606e-06 4.953998e-06  -1.804927 7.127756e-02
DIR*NP^2  -8.437671e-07 6.070856e-08 -13.898653 0.000000e+00
HC         2.069802e-04 7.425082e-06  27.875808 0.000000e+00
```


Residuals From Final Fit
Residuals
0.003
0.002
0.001
0.0
0
10
20
30
40
Nominal Position, in

Summary: Section 1

Regression is a collection of methods used to concisely describe multivariate data as the sum of a function and a probability distribution.

The basic steps in any regression analysis include:

1. selection of the regression function
   - by plotting and using scientific knowledge
2. estimation of the model parameters
   - usually done using ‘least squares’
3. and model validation
   - via graphical residual analysis