SOLUTION OF REGRESSION MODELS

Adequacy of Regression Models

2007 *Jamie Trahan , Autar Kaw*
University of South Florida
United States of America
kaw@eng.usf.edu
http://numericalmethods.eng.usf.edu

Introduction

This worksheet allows you to determine whether a straight-line regression model adequately describes *n* data points (x_1,y_1), (x_2, y_2), (x_3,y_3),....,(x_n,y_n). Four different checks are used to find if the model is adequate. Our discussion, although limited to s line models, is applicable to any regression model.

Adequacy Check #1: Plot of straight-line regression model versus data to visually ins how well the data fits the line.
Adequacy Check #2: Calculation of the coefficient of determination, r^2. This value quantifies the percentage of the original uncertainty in the data that is explained by th straight line model.
Adequacy Check #3: Determine if the residuals as a function of *x* show nonlinearity. is an indication that the model is not adequate.
Adequacy Check #4: Determine if 95% of the values of scaled residuals are within [- If so, this is an indication that the model may be adequate.

Please note that the above checks are not a complete test of the adequacy of regressio models. Other tests include testing if indeed *y* is dependent on *x*, confidence intervals constants of the model, etc.

To learn more about the quality of a fitted linear regression model, see the worksheet the **Adequacy of Regression models**.

Section 1: Input Data

Below are the input parameters to begin the simulation. This is the only section that r user input.
NOTE: The origin has been set to 1 to redefine that starting index of all arrays. The u SHOULD NOT change this value.

ORIGIN:= 1

- Array of *x* values, **X**

NOTE: The user has the option of choosing his or her own *X* and *Y* array (e.g. X = (1,7,13,19,25), Y := (1,49,169,361,625)). We are instead showing a large data se generated within a loop to better illustrate model adequacy. Click on the table a use the scroll bar to see all data points genera

$$X := \left| \begin{array}{l} X_{50} \leftarrow 0 \\ \text{for } i \in 1..50 \\ \quad X_i \leftarrow \frac{i}{2} \\ X \end{array} \right.$$

X =

	1
1	0.5
2	1
3	1.5
4	2
5	2.5
6	3
7	3.5
8	4
9	4.5
10	5
11	5.5
12	6
13	6.5
14	7
15	7.5

- Array of *y* values, **Y**

$$Y := \left| \begin{array}{l} Y_{50} \leftarrow 0 \\ \text{for } i \in 1..50 \\ \quad Y_i \leftarrow (X_i)^2 \\ Y \end{array} \right.$$

Y =

	1
1	0.25
2	1
3	2.25
4	4
5	6.25
6	9
7	12.25
8	16
9	20.25
10	25
11	30.25
12	36
13	42.25
14	49
15	56.25
16	64

Section 2: Finding the straight-line model

We will use Mathcad's *Line* command to fit the data to a straight line.

$n := \text{rows}(X)$

$y := \text{line}(X, Y)$

$a := y_1$

$b := y_2$

The straight-line regression model is

$$f(x) := a + b \cdot x$$

where

$$a = -110.5$$

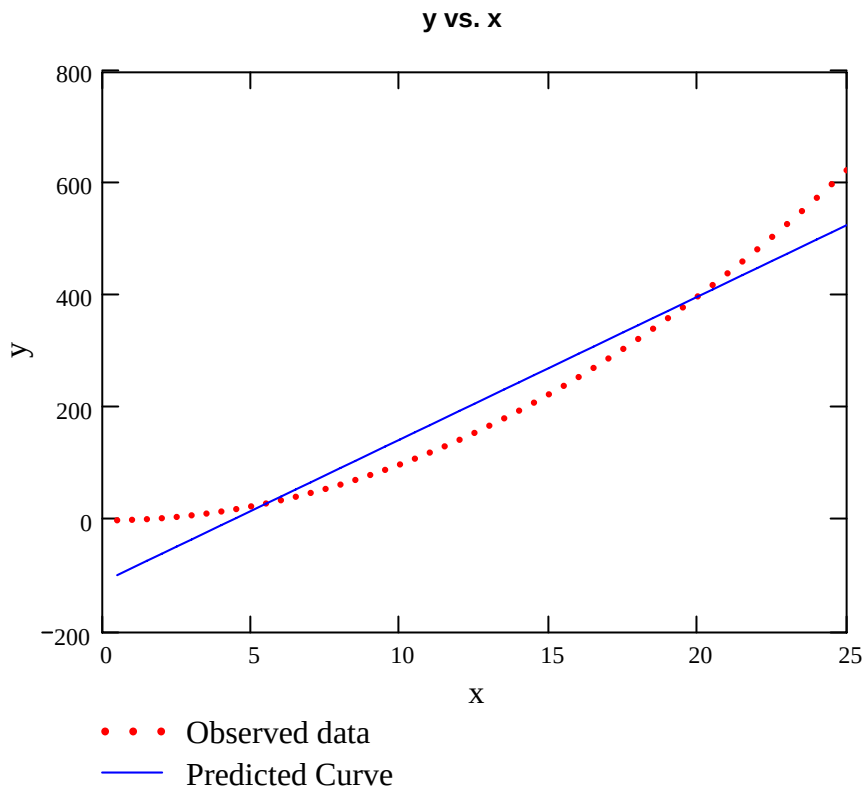
and

$$b = 25.5$$

Section 3: Checking for the adequacy of the model

- **Adequacy Check #1**

  Below, the linear regression model is plotted versus the data points. See if the straight-line regression model visually explains the data.



- **Adequacy Check #2**

In this section we will calculate the coefficient of determination, r^2.

$$r^2 := \frac{S_t - S_r}{S_t}$$

where
S_r = the sum of the squares of the residuals (a value that quantifies the spread around the regression line)
and
S_t = the sum of the squares of deviation from the mean (a value that measures the spread between the data and its mean)

This value describes the proportion of variation in the response data that is explained by the regression model. When all the points in a data set lie on the regression model, the largest possible value of $r^2 = 1$ is obtained, while a minimum possible value of $r^2 = 0$ is obtained when there is only one data point or if the straight line regression model is a constant line.
Note: Please see the **adequacy of regression models** worksheet for limitations in the use of r^2.

Calculation of r^2:

Sum of the difference between observed values and average values, S_t:

$$St := \sum_{i=1}^{n} \left[\left(Y_i - \text{mean}(Y) \right)^2 \right]$$

$$St = 1.801 \times 10^6$$

Sum of the square of the residuals, S_r:

$$Sr := \sum_{i=1}^{n} \left(Y_i - f\left(X_i\right) \right)^2$$

$$Sr = 1.083 \times 10^5$$

Coefficient of determination, r^2:

$$r2 := \frac{St - Sr}{St}$$

r2 = 0.94

- **Adequacy Check #3**

In this section, the residuals, which are the differences between the observed valu and predicted values (y_i - a_0 - a_1x_i), are found and then plotted as a function of x t check for increasing variance, outliers, or nonlinearity.

Calculating the residuals:

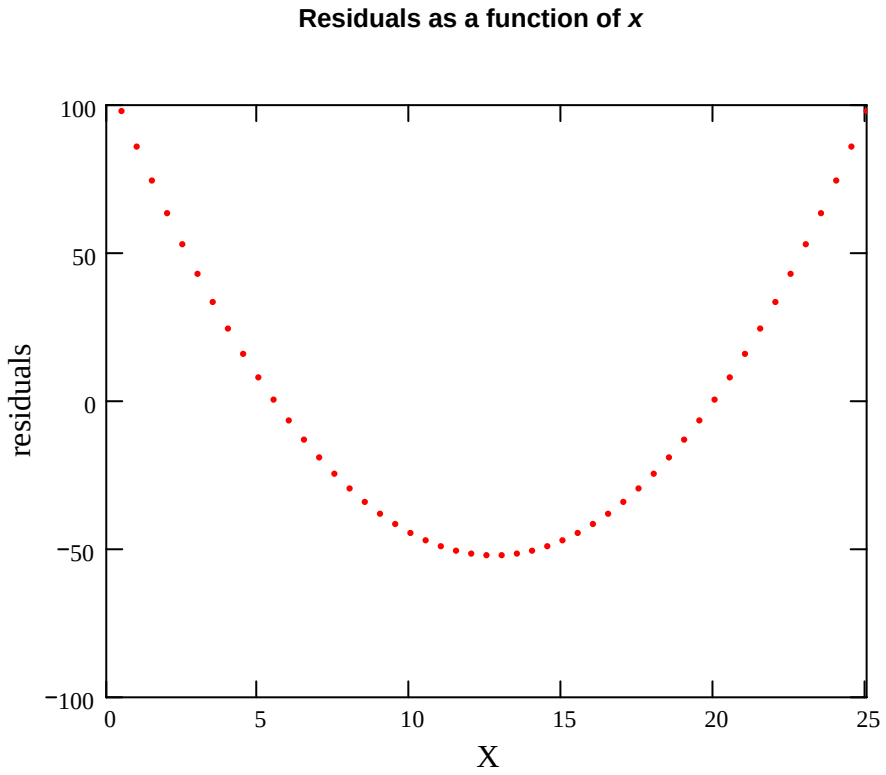
$$\text{residuals} := \left| \begin{array}{l} \text{residuals}_n \leftarrow 0 \\ \text{for } i \in 1 .. n \\ \quad \text{residuals}_i \leftarrow Y_i - f(X_i) \\ \text{residuals} \end{array} \right.$$

The residuals are:

residuals =

	1
1	98
2	86
3	74.5
4	63.5
5	53
6	43
7	33.5
8	24.5
9	16
10	8
11	0.5
12	-6.5
13	-13
14	-19
15	-24.5
16	-29.5

Plotting the residuals:



- **Adequacy Check #4**

In this section, the scaled residuals *SR* (ratio between residual and the standard error of estimate) are calculated. *SR* is given by

$$SR := \frac{y_i - f(x_i)}{\sqrt{\frac{Sr}{n - m}}}$$

where *f(x)* is the regression function, *n* is the number of data points and *m* is the number of degrees of freedom lost (i.e. the number of constants in the model. For a straight-line model, *m* = 2).

Calculation of SR:

$m := 2$

$$SR := \left| \begin{array}{l} SR_n \leftarrow 0 \\ \text{for } i \in 1 .. n \\ \quad SR_i \leftarrow \dfrac{residuals_i}{\sqrt{\dfrac{Sr}{n - m}}} \\ SR \end{array} \right.$$

Calculating the percent within range:

$$count := \left| \begin{array}{l} count \leftarrow 0 \\ \text{for } i \in 1 .. n \\ \quad count \leftarrow count + 1 \quad \text{if } -2 \le SR_i \le 2 \\ count \end{array} \right.$$

$$percent\_within\_range := \frac{count}{n} \cdot 100$$

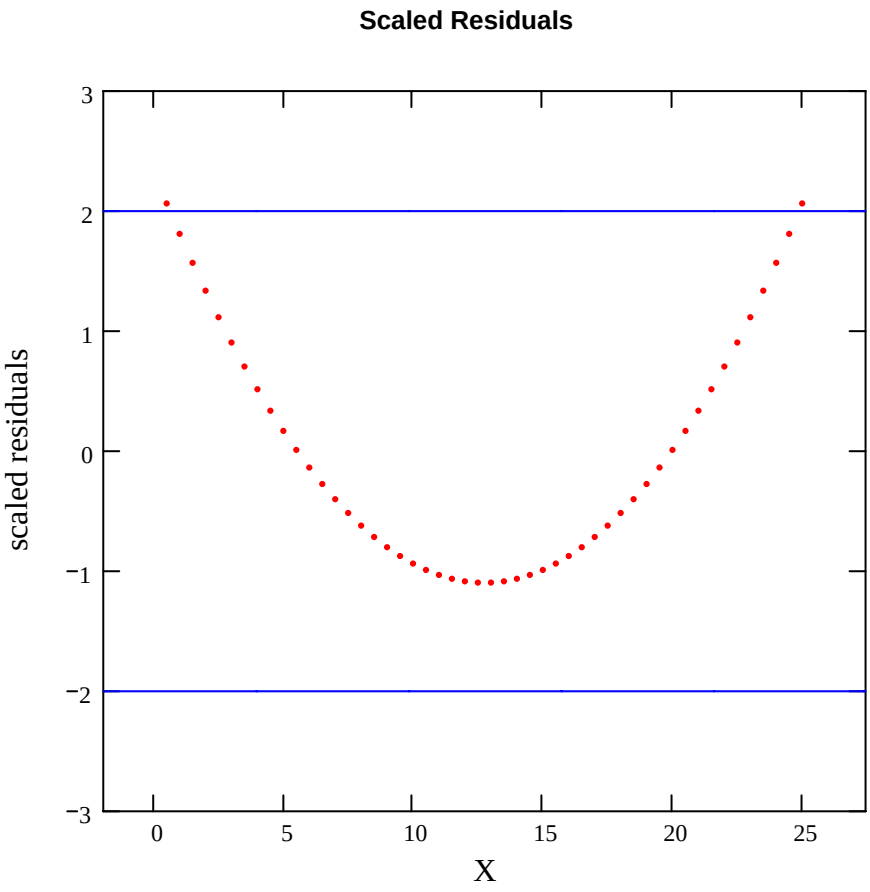
$percent\_within\_range = 96$

$$check := \left| \begin{array}{l} \text{"At least 95\% of SR values are in the [-2,2] range"} \quad \text{if } \dfrac{count}{n} \ge .95 \\ \text{"Less than 95\% of SR values fall in the [-2,2] range"} \quad \text{if } \dfrac{count}{n} < .95 \end{array} \right.$$

$check = \text{"At least 95\% of SR values are in the [-2,2] range"}$

Plotting the scaled residuals:

$g(x) := -2 \qquad h(x) := 2$



References

[1]Autar Kaw, *Holistic Numerical Methods Institute, http://numericalmethods.eng.usf.edu*, See
Adequacy of Regression models
How does Linear Regression work?

Conclusion

Using Mathcad we are able to check for the adequacy of a straight-line regression model.

Question 1: Given data

$i := 1 .. 6$

$Q_i :=$

1
5
3
2
10
20

$Z_i :=$

1
25
9
4
100
400

show if regressing the data to $y = a_0 + a_1 x$ is adequate.

Question 2: Theoretical considerations assume that the rate of flow from a fire hose proportional to some power of the nozzle pressure. However a scientist believes that simpler linear regression model is adequate. Determine whether the linear model is adequate.

$i := 1 .. 7$

Flow rate, F (*gallons/min*) Pressure, p (*psi*)

$F_i :=$

91
120
127
190
240
310
409

$p_i :=$

9
15
23
40
61
72
90

Question 3: Given (x_1,y_1), (x_2, y_2), (x_3,y_3),....,(x_n,y_n), the coefficient of determination, r^2 is zero if the straight line regression model turns out to be a constant line.

Prove that the constant line is the average of the y-values, that is, $\frac{\sum_{i=1}^{n} y_i}{n}$

ɪ:
t
ɪd