Adequacy of Regression Models

 Autar Kaw, Jamie Trahan

University of South Florida

United States of America

kaw@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 straight line models, is applicable to any regression model.

Adequacy Check #1: Plot of straight-line regression model vs. data to visually inspect 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 the straight line model.
Adequacy Check #3: Determine if the residuals as a function of x show nonlinearity. This is an indication that the model is not adequate.
Adequacy Check #4: Determine if 95% of the values of scaled residuals are within [-2,2]. 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 regression models. Other tests include testing if indeed y is dependent on x, confidence intervals of the constants of the model, etc.

To learn more about the quality of a fitted linear regression model, see the worksheet on the adequacy of regression models.

Section 1: Input Data

Below are the input parameters to begin the simulation. This is the only section that requires user input.

Input Parameters:

X = array of x values

Y = array of *y* values
n = number of data points
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 set generated within a loop to better illustrate model adequacy.

```
> restart;
```

```
> X :=array(1..50):
  for i from 1 by 1 to 50 do
       X[i]:=evalf(i/2):
  end do:
  print(X);
```

[0.5000000000, 1., 1.500000000, 2., 2.500000000, 3., 3.500000000, 4., 4.500000000, 5., 5.500000000, 6., 6.500000000, 7., 7.500000000, 8., 8.500000000, 9., 9.500000000, 10., 10.50000000, 11., 11.50000000, 12., 12.50000000, 13., 13.50000000, 14., 14.50000000, 15., 15.50000000, 16., 16.50000000, 17., 17.50000000, 18., 18.50000000, 19., 19.50000000, 20., 20.50000000, 21., 21.50000000, 22., 22.50000000, 23., 23.50000000, 24., 24.50000000, 25.] **(2.1)**

```
> Y:=array(1..50):
  for i from 1 by 1 to 50 do
       Y[i]:=evalf((X[i])^2);
  end do:
  print(Y);
```

[0.2500000000, 1., 2.250000000, 4., 6.250000000, 9., 12.25000000, 16., 20.25000000, 25., 30.25000000, 36., 42.25000000, 49., 56.25000000, 64., 72.25000000, 81., 90.25000000, 100., 110.2500000, 121., 132.2500000, 144., 156.2500000, 169., 182.2500000, 196., 210.2500000, 225., 240.2500000, 256., 272.2500000, 289., 306.2500000, 324., 342.2500000, 361., 380.2500000, 400., 420.2500000, 441., 462.2500000, 484., 506.2500000, 529., 552.2500000, 576., 600.2500000, 625.] **(2.2)**

```
> n:=50;
```

$n := 50$ **(2.3)**

Section 2: Finding the straight-line model

We will use Maple's *Fit* command to fit the data to a straight line.

```
> with(Statistics):
  y:=Fit(a*x+b,X,Y,x):
  f:=unapply(y,x):
```

```
print(`The straight-line regression model is y = `, y );
```

```
>
```

The straight-line regression model is y = , $-110.499999999999815 + 25.4999999999999858\,x$ **(3.1)**

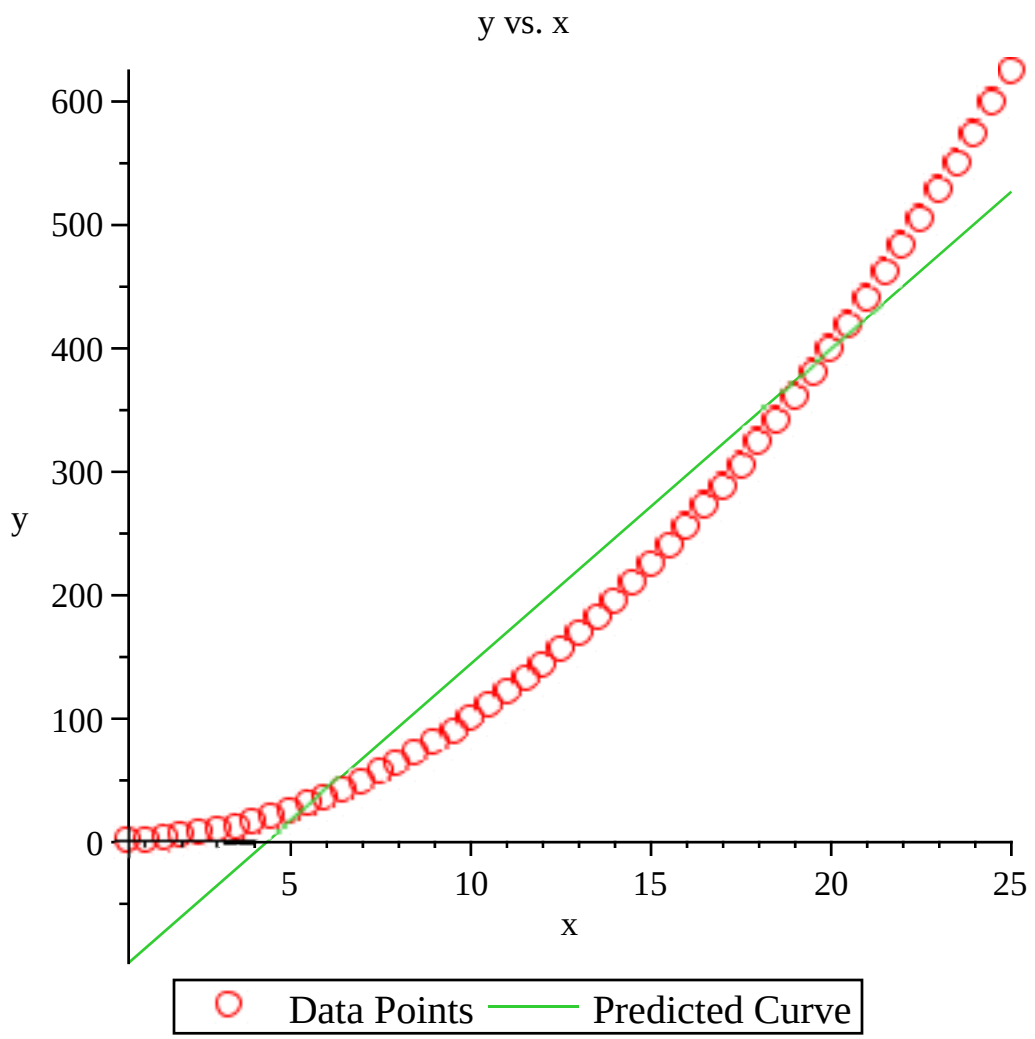
Section 3: Checking for adequacy of the model

• ***Adequacy Check #1***

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

```
> observed:=[seq([X[i],Y[i]],i=1..n)]:
  predicted:=f(x):
  plot([observed,predicted],x=X[1]..X[n],style=[POINT,LINE],
  labels=["x","y"],symbol=CIRCLE,symbolsize=20,title="y vs. x",
  legend=["Data Points","Predicted Curve"]);
```

```
>
```



>

• ***Adequacy Check #2***

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

$r^2 = \frac{St - Sr}{St}$

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 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:

```
> with(Statistics):
  St:=0:
  for i from 1 by 1 to n do
       St:=St+(Y[i]-Mean(Y))^2;
  end do:
  St;
```

$$1.800972033\ 10^6 \quad \textbf{(4.1)}$$

Sum of the square of the residuals, S_r:

```
> Sr:=0:
  for i from 1 by 1 to n do
       Sr:=Sr+(Y[i]-f(X[i]))^2;
  end do:
  Sr;
```

$$1.082900000\ 10^5 \quad \textbf{(4.2)}$$

Coefficient of determination, r^2:

```
> r2:=(St-Sr)/St;
```

$$r2 := 0.9398713595 \quad \textbf{(4.3)}$$

• ***Adequacy Check #3***

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

Calculating the residuals:

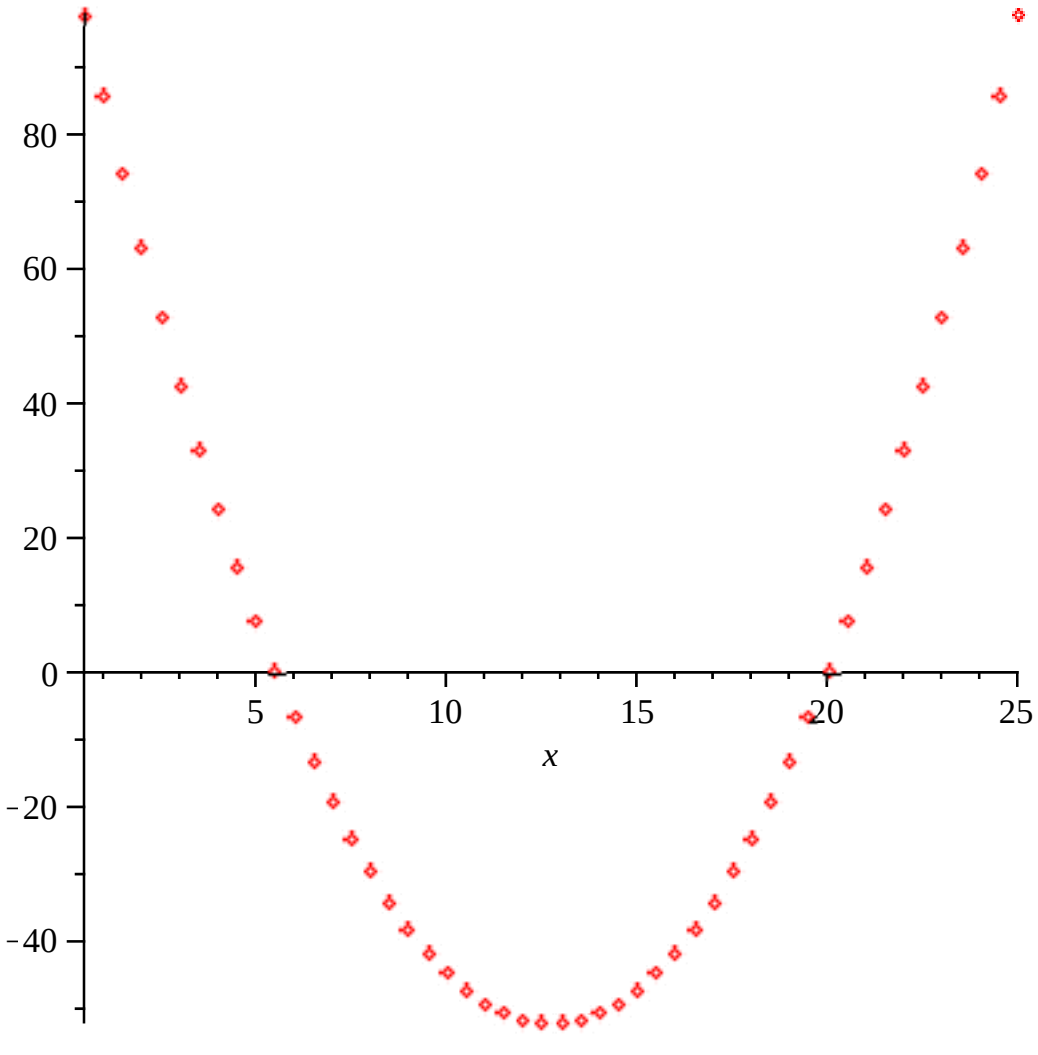
```
> residuals:=array(1..n):
  for i from 1 by 1 to n do
       residuals[i]:=Y[i]-f(X[i]);
  end do:
  print(`The residuals are `,residuals);
```

The residuals are , [98.00000000, 86.00000000, 74.50000000, 63.50000000, 53.00000000, 43.00000000, 33.50000000, 24.5000000, 16.00000000, 8.0000000, 0.50000000, −6.5000000, −13.00000000, −19.0000000, −24.50000000, −29.5000000, −34.00000000, −38.0000000, −41.50000000, −44.5000000, −47.0000000, −49.0000000, −50.5000000, −51.5000000, −52.0000000, −52.0000000, −51.5000000, −50.5000000, −49.0000000, −47.0000000, −44.5000000, **(4.4)**

−41.5000000, −38.0000000, −34.0000000, −29.5000000, −24.5000000, −19.0000000, −13.0000000, −6.5000000, 0.5000000, 8.0000000, 16.0000000, 24.5000000, 33.5000000, 43.0000000, 53.0000000, 63.5000000, 74.5000000, 86.0000000, 98.0000000]

Plotting the residuals.

```
> residual:=[seq([X[i],residuals[i]],i=1..n)]:
  plot(residual,x=X[1]..X[n],style=point);
```



• *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

$$SR = \frac{y[i] - f(x[i])}{sqrt(Sr/(n-m)}$$

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:=array(1..n):
  for i from 1 by 1 to n do
       SR[i]:=residuals[i]/(sqrt(Sr/(n-m)));
  end do:
```

```
> print(SR);
```

[2.063253153, 1.810609910, 1.568493468, 1.336903829, 1.115840991, 0.9053049550, 0.7052957207, 0.5158132883, 0.3368576577, 0.1684288288, 0.01052680180, −.1368484234, −.2736968468, −.4000184685, −.5158132883, −.6210813063, −.7158225225, −.8000369369, −.8737245496, −.9368853604, −.9895193694, −1.031626577, −1.063206982, −1.084260586, −1.094787387, −1.094787387, −1.084260586, −1.063206982, −1.031626577, −.9895193694, −.9368853604, −.8737245496, −.8000369369, −.7158225225, −.6210813063, −.5158132883, −.4000184685, −.2736968468, −.1368484234, 0.01052680180, 0.1684288288, 0.3368576577, 0.5158132883, 0.7052957207, 0.9053049550, 1.115840991, 1.336903829, 1.568493468, 1.810609910, 2.063253153] **(4.5)**

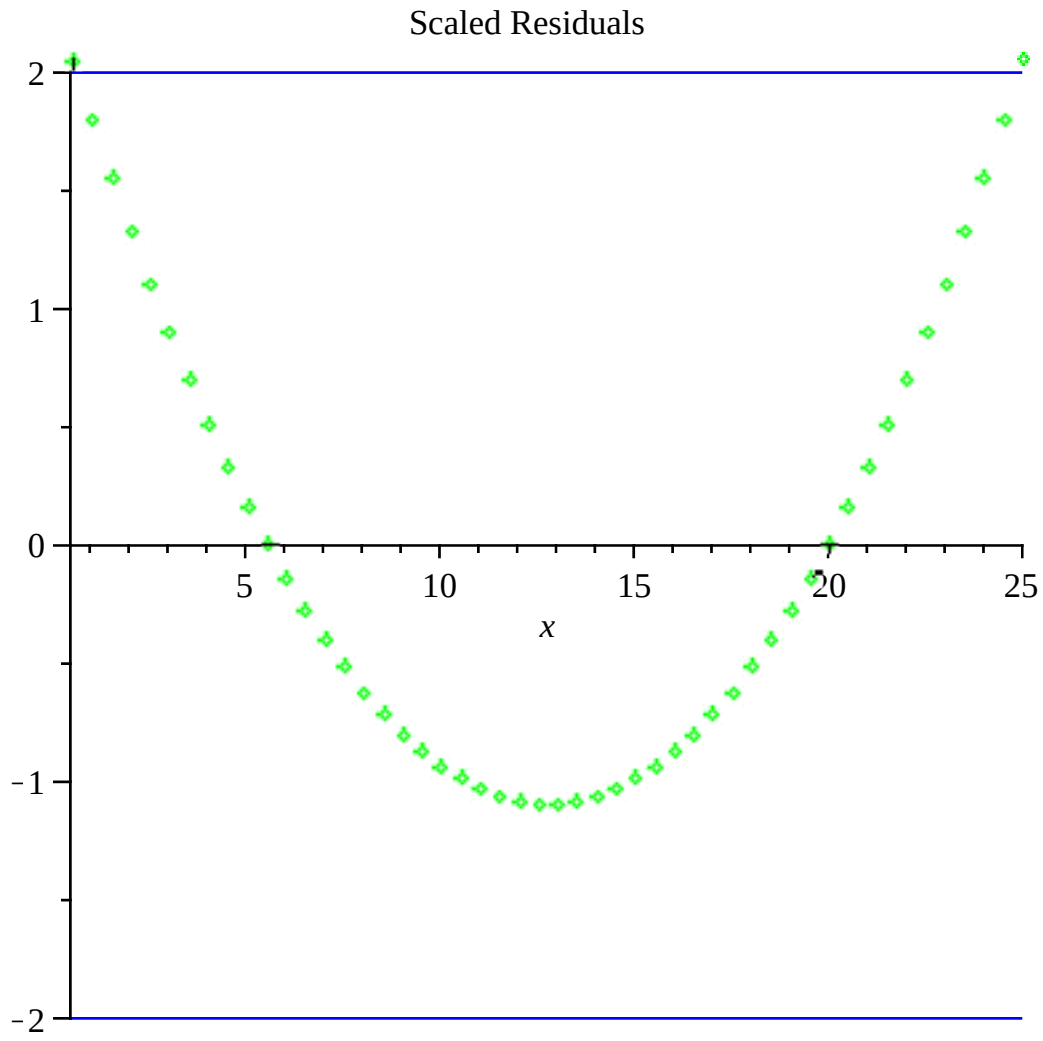
Calculating the percent within range:

```
> count:=0:
  for i from 1 by 1 to n do
       if -2<=SR[i] and SR[i]<=2 then
            count:=count+1;
       end if:
  end do:
  if (count/n)>=0.95 then
       print((count/n)*100,`% of SR values fall between -2 and
  2, therefore at least 95% of SR values are in the [-2,2]
  range`);
  else
       print((count/n)*100, `% of SR values fall between -2 and
  2, therefore at least 95% of SR values are not in the [-2,2]
  range`);
  end if;
```

96, *% of SR values fall between -2 and 2, therefore at least 95% of SR values are in the [-2, 2] range* **(4.6)**

Plotting the scaled residuals:

```
> scaledresid:=[seq([X[i],SR[i]],i=1..n)]:
  plot([2,-2,scaledresid],x=X[1]..X[n],style=[line,line,point],
  color=[blue,blue,green],title="Scaled Residuals");
```



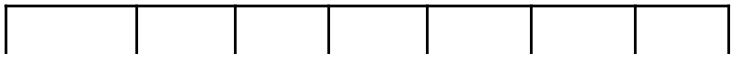
References

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

Conclusion

Using Maple we are able to check for the adequacy of a linear regression model.

Question 1: Given data



x	1	5	3	2	10	20
y	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 is proportional to some power of the nozzle pressure. However a scientist believes that the simpler linear regression model is adequate. Determine whether the linear model is adequate.

Flow rate, F $(gallons/min)$	91	120	127	190	24\	310	409
Pressure, p (psi)	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, r2 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, $\frac{\sum y[i],\ i = 1..n}{n}$.