R Introduction to Linear Regression

jdt

11/26/2020

Contents

Regression Theory **1**
- Least Squares Solution . . . 2
- Inference . . . 4
- Analysis of Variance for Regression . . . 5

R code **6**
- Non matrix approach . . . 6

Matrix Approach **8**
- Least Squares Solution . . . 8
- Geometry of Ordinary Least Squares . . . 10

R Code with matrix **11**
- my.regress function . . . 13
- my.regress.plot Function . . . 14

R packages for Regression **19**

Assignment **21**
- Non matrix Notation . . . 21
- Matrix Notation . . . 22

SAS Program **23**
- Code . . . 23
- Output . . . 23

Regression Theory

In this document, we consider a modeling problem whereby we are interested in determining the relationship between two (or more) random variables from a single population of interest. These models are called *regression models.*

Before beginning with the simplest of these models, the linear model, given by

$$\mu_y = E(Y) = \beta_0 + \beta_1 X \tag{1}$$

where the mean, μ_y, of the dependent variable of interest, Y, can be written as a linear function of a independent variable X as determine by two parameters, the slope, β_1, and y-intercept, β_0.

There are two main purposes or objectives for creating a regression model.

- The first is **prediction** where the objective is to predict the expected value of Y at a specified value of X. For example, a university might wish to predict the expected enrollment of a freshman class using an early indicator variable, X, say, deposits at the April 1.
- The second purpose, and in my opinion, the more important use of regression models is **control**. Where the independent variable X acts as a controlling variable for unexplained variation in the mean population response or independent variable, Y. For example, suppose that Y is an individual's systolic blood pressure for which high values are indicative of poor health-related outcomes. So how does one reduce this blood pressure value, thereby, hopefully improving one's health outcomes? Studies have shown that sodium intake and exercise level can effect blood pressure. If so, can you control your sodium intake or exercise level in hopes of controlling your blood pressure? If so, you have a regression model and a good reason for creating one.

Least Squares Solution

The mathematical concept called least squares is used in regression. It is also used when finding an estimate for μ when $X \sim N(\mu, \sigma^2)$ as given below

Least Squares for the Mean

The problem (mathematical) is to find the constant, c, that minimizes the following:

$$Q(c) = \sum_{i=1}^{n} (y_i - c)^2 \tag{2}$$

where the data are $y_1, y_2, \ldots, y_n$. Since $Q(c)$ is a univariate function of c, the solution is a simple calculus problem given as the solution of

$$\begin{aligned} \frac{dQ(c)}{dc} &= 0 \\ 2\sum_{i=1}^{n}(y_i - c) &= 0 \\ \sum_{i=1}^{n} y_i &= nc \\ c = \frac{1}{n}\sum_{i=1}^{n} y_i &= \bar{y}. \end{aligned}$$

Hence, the arithmetic mean is the value that minimizes the sum of squared "centered" values of the n y_i's. The above solution is found using calculus, yet the solution can be verified as follows. Let the constant c be

any real number and consider

$$\begin{aligned} Q(c) &= \sum_{i=1}^{n}(y_i - c)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - c)^2 \\ &= \sum_{i=1}^{n}[(y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - c) + (\bar{y} - c)^2] \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - c)^2 \\ &> \sum_{i=1}^{n}(y_i - \bar{y})^2 \end{aligned}$$

since $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$ and $n(\bar{y} - c)^2 \geq 0$.

This approach is taken when finding the least squared error solution to the best linear equation when one has 2 variables.

Suppose that one has two random variables X and Y for which one observes n pairs, denoted by (x_i, y_i), $i = 1, 2, \ldots, n$. The objective of the least squares problem is to determine, $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the linear equation given by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ minimizes[1]

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2 = \sum_{i=1}^{n}\epsilon_i^2. \tag{3}$$

where $\epsilon_i = (y_i - (\beta_0 + \beta_1 x_i))$ is called the *population residual or error* for the observation (x_i, y_i). The solution can be determined by taking the partial derivatives of $Q(\beta_0, \beta_1)$ with respect to both β_0 and β_1. That is,

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}[y_i - \beta_0 - \beta_1 x_i]$$

and

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n}[y_i - \beta_0 - \beta_1 x_i]x_i.$$

By setting these equations equal to zero, one obtains

$$\begin{aligned} n\beta_0 + \sum x_i \beta_1 &= \sum y_i \\ \sum x_i \beta_0 + \sum x_i^2 \beta_1 &= \sum x_i y_i. \end{aligned} \tag{4}$$

Equation(4) is called **the normal equations** for the linear least squares problem. From which, the unique solution is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{5}$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{6}$$

[1] In this chapter I will present the material in much the same way as I would if teaching STAT 2381 (albeit a little faster). I will not use matrix notation in this chapter, whereas I will in the remainder part of the course notes.

where

$$\begin{aligned}
\bar{y} &= \sum_{i=1}^{n} y_i/n \qquad \bar{x} = \sum_{i=1}^{n} x_i/n \\
SS_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} \\
SS_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = (n-1)s_x^2 \\
SS_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = (n-1)s_y^2
\end{aligned}$$

The estimated residual, $\hat{e}_i = y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i)$, is the vertical distance that the observed y_i is from the least squares line $(\hat{\beta}_0 + \hat{\beta}_1 x)$ at $x = x_i$. The *residual sum of squares* or *sum of squares due to the error* is

$$SS_E = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum \hat{e}_i^2 = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}.$$

The predicted value (line at $x = x^*$) can be written as[2][3]

$$\hat{\mu}_{y|x^*} = \hat{y}^* = \bar{y} + \hat{\beta}_1(x^* - \bar{x})$$

Inference

The procedure for computing the least squares estimates for β_0 and β_1 is a mathematical optimization problem. In order to use this method as a statistical problem one must make additional distributional assumptions concerning the response or dependent variable y. These assumptions are;

- The observed dependent data $y_1, y_2, \ldots, y_n$, are a sample from a population where $Y \sim N(\mu_{y_i}, \sigma_y^2)$. This is called the **normality assumption**.
- $E(y_i) = \mu_{y_i} = \beta_0 + \beta_1 x_i$. The expected value for y_i is a linear function of x_i. This is called the **linear assumption**.
- σ_y^2 does not depend upon the value of x_i. This is called the **homogeneity of variance assumption**.
- The data $y_1, y_2, \ldots, y_n$ are independent. This is called the **independence assumption**.

The above assumptions allow one to determine the standard errors for the statistical estimates of the population parameters of interest; β_0, β_1, and $E(y \mid x = x^*) = \mu_{y|x^*}$.

- The slope (β_1)

  $$\hat{\beta}_1 = SS_{xy}/SS_{xx}$$

  and

  $$\hat{\sigma}_{\hat{\beta}_1} = \hat{\sigma}/\sqrt{SS_{xx}}$$

  where

  $$\hat{\sigma} = \sqrt{SSE/(n-2)} = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}.$$

[2]Note if the estimated slope $\hat{\beta}_1$ is zero then the predicted value for every y_i is $\bar{y}$. Which indicates that the variable X was not needed in the regression model for Y.

[3]Every linear least squares regression line passes through the point $(\bar{x}, \bar{y})$.

- The line ($\mu_{y|x^*}$)

  $$\hat{\mu}_{y|x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* = \bar{y} + \hat{\beta}_1(x^* - \bar{x})$$

  and

  $$\hat{\sigma}_{\hat{\mu}_{y|x^*}} = \hat{\sigma}\sqrt{1/n + (x^* - \bar{x})^2/SS_{xx}}.$$

- The y-intercept (β_0)

  $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

  and

  $$\hat{\sigma}_{\hat{\beta}_0} = \hat{\sigma}\sqrt{1/n + \bar{x}^2/SS_{xx}}.$$

The $(1-\gamma)100\%$ Confidence intervals are of the form

$$\text{(estimate)} \pm t_{\gamma/2}(df = (n-2)) \times \text{(standard error of estimate)}.$$

where $t_{\gamma/2}(df = (n-2))$ is the critical point from a t-distribution with $df = n - 2$.

Analysis of Variance for Regression

The regression results are often presented as an analysis of variance table or ANOVA table. The basic idea is to describe how much of the variability found in the dependent variable y can be explained by the presence of the linear equation ($\beta_1 \neq 0$) versus having a line $y = \bar{y}$ ($\beta_1 = 0$).

The total adjusted (corrected for β_0) sum of squares in the dependent variable y is given by $SS_{CT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$. This corrected sum of squares can be written as $SS_{CT} = SS_M + SS_E$ where

$$SS_M = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

and

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

SS_M is the sum of squares due to the model and SS_E is the sum of squares due to the error (*Residual sum of squares*).

The above follows from

$$\begin{aligned}
\sum(y_i - \hat{y}_i)^2 &= \sum(y_i - \bar{y} + \bar{y} - \hat{y}_i)^2 \\
&= \sum(y_i - \bar{y})^2 + (\hat{y}_i - \bar{y})^2 - 2(y_i - \bar{y})(\hat{y}_i - \bar{y}) \\
&= \sum(y_i - \bar{y})^2 + \sum(\hat{y}_i - \bar{y})^2 - 2\sum(y_i - \bar{y})(\hat{y}_i - \bar{y}) \\
&= \sum(y_i - \bar{y})^2 - \sum(\hat{y}_i - \bar{y})^2
\end{aligned}$$

where

$$\begin{aligned}
-2\sum(y_i - \bar{y})(\hat{y}_i - \bar{y}) &= -2\sum(y_i - \bar{y})\hat{\beta}_1(x_i - \bar{x}) \\
&= -2\hat{\beta}_1\sum(y_i - \bar{y})(x_i - \bar{x}) \\
&= -2\hat{\beta}_1^2\sum(x_i - \bar{x})^2 \\
&= -2\sum(\hat{y}_i - \bar{y})^2.
\end{aligned}$$

If the slope of the line is nonzero then one would expect that a sizeable amount of the variability in y as specified by SS_{CT} would be attributable to SS_M. One way of measuring this is to compute $R^2 = SS_M/SS_{CT} = SS_M/(SS_M + SS_E)$. R^2 is a number between 0 and 1 which is usually expressed as a percentage. **Since SS_{CT} has been adjusted for $\hat{\beta}_0$, R^2 is the variability explained by the model relative to what can be explained by $\bar{y}$.** The closer the value is to 1 or 100% means that the amount of variability found in SS_{CT} is nearly explained by the model (or in this case having $\hat{\beta}_1$ be nonzero). On the other hand if R^2 is close to zero then very little of the variability in the data is explained by the model as opposed to just using $\hat{\mu}_y = \bar{y}$, which means that one doesn't need x in order to explain variability in y.

ANOVA Table

The analysis of variance table is given by

Source	Degrees of Freedom	Sum of Squares	Mean Square
due to $\beta_1 \mid \beta_0$	1	$SS_M = \sum(\hat{y}_i - \bar{y})^2$	$MS_M = SS_M/1$
Residual	n-2	$SS_E = \sum(y_i - \hat{y}_i)^2$	$MS_E = SS_E/(n-2)$
Corrected Total	n-1	$SS_{CT} = \sum(y_i - \bar{y})^2$	
due to β_0	1	Correction factor =$n\bar{y}^2$	
Total	n	$SS_T = \sum y_i^2$	

R code

Enter the Data

```
x=c(15.6,26.8,37.8,36.4,35.5,18.6,15.3,7.9,0)
y=c(5.2,6.1,8.7,8.5,8.8,4.9,4.5,2.5,1.1)
```

Non matrix approach

```
n<-length(x)
ssxy<- sum((x-mean(x))*(y-mean(y)))
ssxx<- sum((x-mean(x))*(x-mean(x)))
ssyy<- sum((y-mean(y))*(y-mean(y)))
sx<-ssxx/(n-1)
sy<-ssyy/(n-1)
ssxy
```

```
## [1] 291.3144
```

```
sx
```

```
## [1] 180.0803
```

```
sy
```

```
## [1] 7.528611
```

```r
b1<-ssxy/ssxx
b0<-mean(y)-b1*mean(x)
b0
```

```
## [1] 1.232354
```

```r
b1
```

```
## [1] 0.2022115
```

```r
y.hat<- b0 + b1*x
y.hat
```

```
## [1] 4.386854 6.651623 8.875949 8.592853 8.410863 4.993488 4.326190 2.829825
## [9] 1.232354
```

```r
y.resid<-y - y.hat
y.resid
```

```
## [1]  0.81314622 -0.55162273 -0.17594938 -0.09285326  0.38913710 -0.09348832
## [7]  0.17380967 -0.32982512 -0.13235417
```

```r
s.se<-sum(y.resid*y.resid)
s.se
```

```
## [1] 1.321754
```

```r
m.se<-s.se/(n-2)
r.mse<-sqrt(m.se)
m.se
```

```
## [1] 0.188822
```

```r
r.mse
```

```
## [1] 0.4345366
```

```r
ub1<- b1 + qt(.975,n-2)*sqrt(m.se/ssxx)
lb1<- b1 - qt(.975,n-2)*sqrt(m.se/ssxx)
ub1
```

```
## [1] 0.2292829
```

```r
lb1
```

```
## [1] 0.1751401
```

```r
u.95<-y.hat + qt(.975,n-2)*r.mse*sqrt(1/n + (x-mean(x))*(x-mean(x))/ssxx)
u.95
```

```
## [1] 4.765280 7.022503 9.433590 9.121098 8.920804 5.345147 4.708141 3.333558
## [9] 1.908724
```

```r
l.95<-y.hat - qt(.975,n-2)*r.mse*sqrt(1/n + (x-mean(x))*(x-mean(x))/ssxx)
l.95
```

```
## [1] 4.0084275 6.2807426 8.3183089 8.0646082 7.9009220 4.6418300 3.9442397
```

```
## [8] 2.3260921 0.5559843
```

```
Y<-cbind(x,y,y.hat,l.95,u.95)
Y
```

```
##          x   y    y.hat      l.95     u.95
##  [1,] 15.6 5.2 4.386854 4.0084275 4.765280
##  [2,] 26.8 6.1 6.651623 6.2807426 7.022503
##  [3,] 37.8 8.7 8.875949 8.3183089 9.433590
##  [4,] 36.4 8.5 8.592853 8.0646082 9.121098
##  [5,] 35.5 8.8 8.410863 7.9009220 8.920804
##  [6,] 18.6 4.9 4.993488 4.6418300 5.345147
##  [7,] 15.3 4.5 4.326190 3.9442397 4.708141
##  [8,]  7.9 2.5 2.829825 2.3260921 3.333558
##  [9,]  0.0 1.1 1.232354 0.5559843 1.908724
```

Matrix Approach

In this section the above methods for the linear least squares problem are reproduced using matrix notation.

The simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{7}$$

for $i = 1, 2, \ldots, n$ where ϵ_i represents the unobserved error (or distance) that the observed data value y_i is from its mean $\mu_{y|x_i} = \beta_0 + \beta_1 x_i$ when $x = x_i$. This model can be written as

$$\mathbf{y} = \mathbf{X}\vec{\beta} + \vec{\epsilon} \tag{8}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \mathbf{j_n} & \mathbf{x} \end{pmatrix} \qquad \vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and

$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

where $\mathbf{j_n}' = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$ and $\mathbf{x}' = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$. As before y denotes a vector rather than $\mathbf{y}$ when the context is clear.

Least Squares Solution

The least squares problem becomes finding $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{pmatrix}'$ that minimizes

$$\begin{aligned} Q(\beta) &= \epsilon'\epsilon \\ &= (y - X\beta)'(y - X\beta) \\ &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &= y'y - 2\beta'(X'y) + \beta'(X'X)\beta \end{aligned}$$

since $Q(\beta)$ is a scalar, one can use the properties of matrix differentiation to find

$$\frac{\partial Q(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0.$$

From which one obtains the normal equations given by

$$X'X\beta = X'y. \tag{9}$$

If $rank[X] = 2$, then the normal equation has a unique solution given by

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{10}$$

Which can be written as

$$\begin{aligned}
\hat{\beta} &= \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \\
&= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\
&= \frac{1}{\sum (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.
\end{aligned}$$

By letting $L = (X'X)^{-1}X'$, it follows that the least squares estimate for β is a linear function of y, given by $\hat{\beta} = Ly$. From which we have that the predicted value for y can be written as $\hat{y} = X\hat{\beta} = XLy = X(X'X)^{-1}X'y = Hy$, where

$$H = X(X'X)^{-1}X', \tag{11}$$

is a $n \times n$ matrix called the ***Hat Matrix***. The computed or estimated residuals are given by $\hat{e} = y - \hat{y} = y - Hy = (I - H)y$ and the residual sum of squares (SS_E) can be written as

$$Q(\hat{\beta}) = \hat{e}'\hat{e} = y'(I - H)'(I - H)y = y'(I - H)y = y'y - \hat{\beta}'X'y.$$

Geometry of Ordinary Least Squares

In the figures given below, let b denote a vector in $\Re^n$ ($b = \mathbf{y}$ in our notation). The objective is to find x [$x = \beta' = (\beta_0, \beta_1)$] so that $\| b - Ax \| = (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta)$ is minimized where $Ax = X\beta$ is a vector on the range space of the matrix A, denoted by range(A)[4] (Figure 1). The minimum is obtained at the orthogonal projection of b onto range(A) (Figure 2).

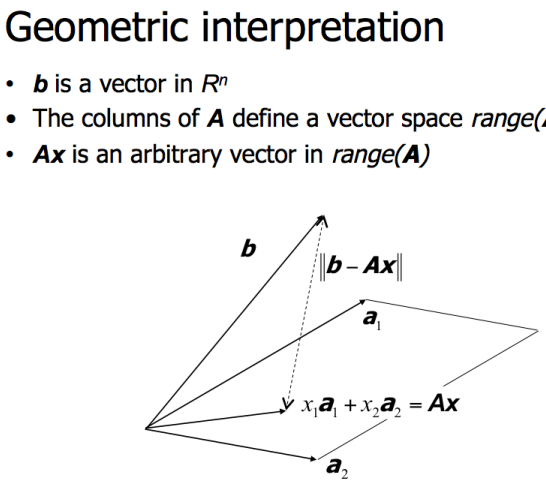


Figure 1: Geometric Interpretation of Least Squares

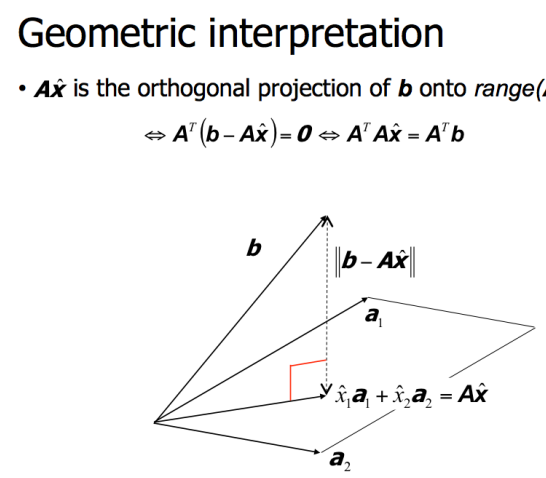


Figure 2: $A\hat{x}$ is the Orthogonal projection

[4] In this chapter the range(A) is a two dimensional plane in $\Re^n$. It is called the column space of X in my notes.

R Code with matrix

```
n<-length(x)
X<-cbind(array(1,c(n,1)),x)
B<-solve(t(X)%*%X)%*%t(X)%*%y
B
```

```
##         [,1]
##   1.2323542
## x 0.2022115
```

```
yhat<-X%*%B
yhat
```

```
##           [,1]
##  [1,] 4.386854
##  [2,] 6.651623
##  [3,] 8.875949
##  [4,] 8.592853
##  [5,] 8.410863
##  [6,] 4.993488
##  [7,] 4.326190
##  [8,] 2.829825
##  [9,] 1.232354
```

```
resid<-y-yhat
resid
```

```
##              [,1]
##  [1,]  0.81314622
##  [2,] -0.55162273
##  [3,] -0.17594938
##  [4,] -0.09285326
##  [5,]  0.38913710
##  [6,] -0.09348832
##  [7,]  0.17380967
##  [8,] -0.32982512
##  [9,] -0.13235417
```

```
sse<-sum(resid^2)
sse
```

```
## [1] 1.321754
```

```
n<-dim(X)[1]
p<-dim(X)[2]
mse<-sse/(n-p)
mse
```

```
## [1] 0.188822
```

```r
cssy<-sum((y-mean(y))^2)
rsquare<-(cssy-sse)/cssy
rsquare
```

```
## [1] 0.9780545
```

```r
stdb<-sqrt(diag(solve(t(X)%*%X))*mse)
stdb
```

```
##                       x
## 0.28603693 0.01144849
```

```r
t<-B/stdb
t
```

```
##         [,1]
##    4.308374
## x 17.662722
```

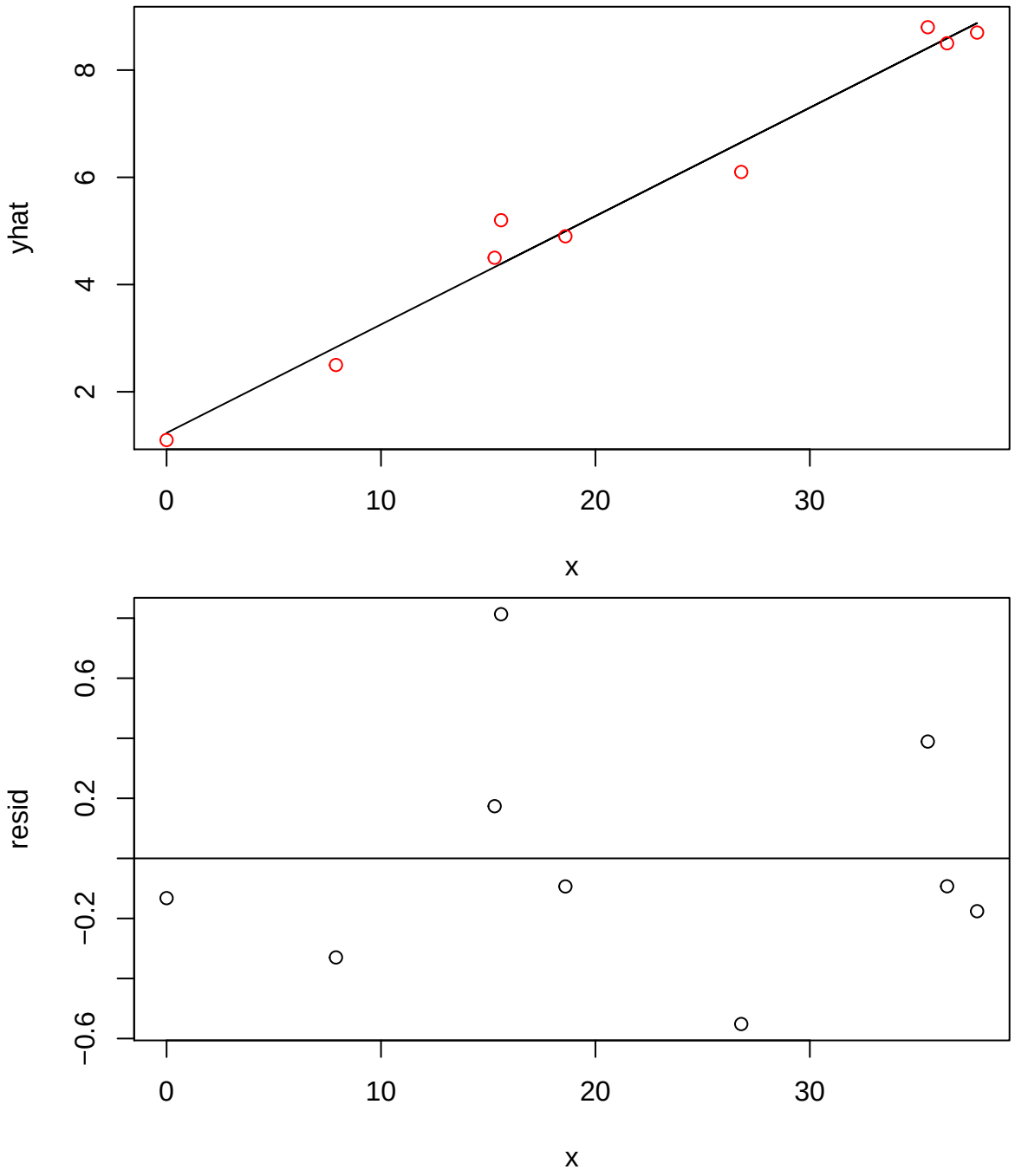
```r
prob.1.sided<-1 - pt(t,n-p)
prob.1.sided
```

```
##            [,1]
##   1.765068e-03
## x 2.298018e-07
```

```r
prob.2.sided<-1 - pf(t*t,1,n-p)
prob.2.sided
```

```
##            [,1]
##   3.530135e-03
## x 4.596036e-07
```

```r
par(mfrow=c(2,1))
plot(x,yhat,type="l")
points(x,y,col="red")
plot(x,resid)
abline(h=0)
```



my.regress function

```
my.regress<-function(x,y){
n<-length(x)
X<-cbind(array(1,c(n,1)),x)
B<-solve(t(X)%*%X)%*%t(X)%*%y
yhat<-X%*%B
resid<-y-yhat
sse<-sum(resid^2)
```

```
n<-dim(X)[1]
p<-dim(X)[2]
mse<-sse/(n-p)
rmse<-sqrt(mse)
cssy<-sum((y-mean(y))^2)
rsquare<-(cssy-sse)/cssy
covb<-(solve(t(X)%*%X))*mse
stdb<-sqrt(diag(covb))
t<-B/stdb
prob1<-1 - pt(t,n-p)
prob2<-1 - pf(t*t,1,n-p)
H<-X%*%solve(t(X)%*%X)%*%t(X)
vresid<-(diag(n)-H)*mse
vpred<-H*mse
h<-diag(H)
lowerrm95<-yhat-qt(.975,n-p)*sqrt(h*mse)
upperrm95<-yhat+qt(.975,n-p)*sqrt(h*mse)
lower95<-yhat-qt(.975,n-p)*sqrt(h*mse+mse)
upper95<-yhat+qt(.975,n-p)*sqrt(h*mse+mse)
list(x=x,y=y,Beta=B,df=n-p,yhat=yhat,resid=resid,sse=sse,
     mse=mse,rmse=rmse,rsquare=rsquare,stdb=stdb,
     covb=covb,t=t,prob1=prob1,prob2=prob2,H=H,vresid=vresid,
     vpred=vpred,h=h,lowerrm95=lowerrm95,uppperrm95=upperrm95,
     lower95=lower95,upper95=upper95)
}
```

my.regress.plot Function

```
my.regress.plot<-function(temp)
{
par(mfrow=c(2,2))
plot(temp$x,temp$yhat,type="l")
points(temp$x,temp$y,col="red")
plot(temp$x,temp$resid)
abline(h=0)
plot(temp$x,temp$h)
qqnorm(temp$resid,main="")
abline(v=0)
qqline(temp$resid)
}

result<-my.regress(x,y)
result
```

```
## $x
## [1] 15.6 26.8 37.8 36.4 35.5 18.6 15.3  7.9  0.0
##
## $y
```

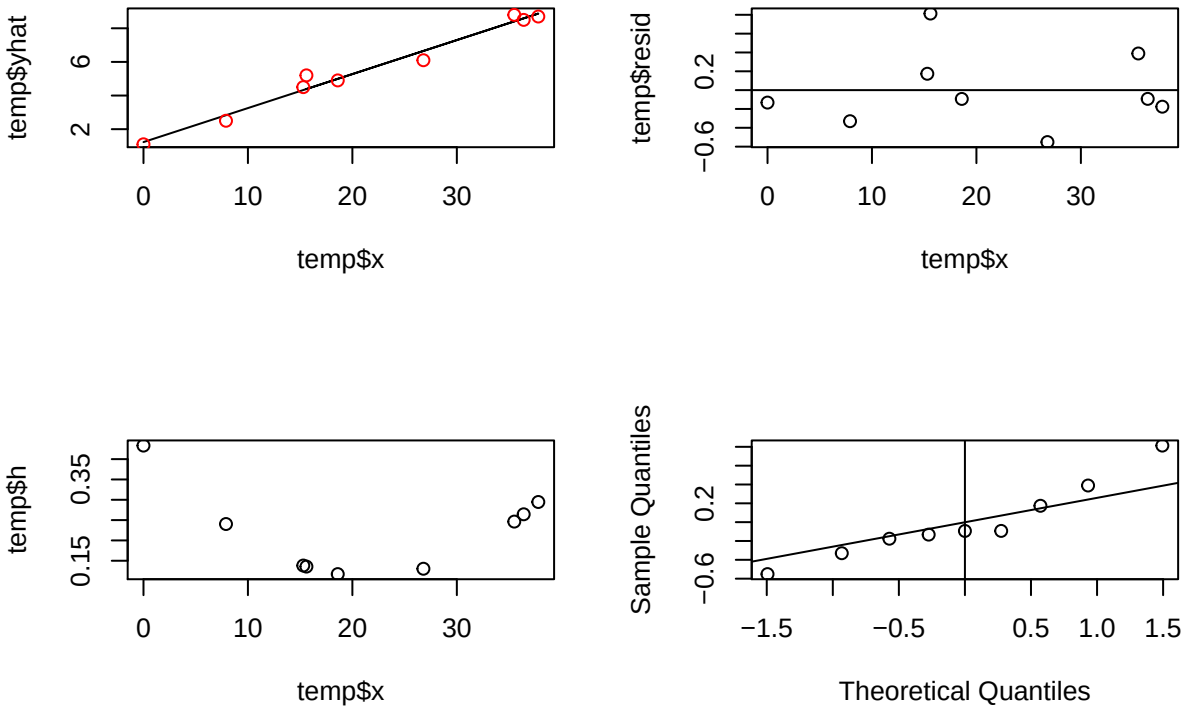
```
## [1] 5.2 6.1 8.7 8.5 8.8 4.9 4.5 2.5 1.1
##
## $Beta
##        [,1]
##   1.2323542
## x 0.2022115
##
## $df
## [1] 7
##
## $yhat
##           [,1]
##  [1,] 4.386854
##  [2,] 6.651623
##  [3,] 8.875949
##  [4,] 8.592853
##  [5,] 8.410863
##  [6,] 4.993488
##  [7,] 4.326190
##  [8,] 2.829825
##  [9,] 1.232354
##
## $resid
##              [,1]
##  [1,]  0.81314622
##  [2,] -0.55162273
##  [3,] -0.17594938
##  [4,] -0.09285326
##  [5,]  0.38913710
##  [6,] -0.09348832
##  [7,]  0.17380967
##  [8,] -0.32982512
##  [9,] -0.13235417
##
## $sse
## [1] 1.321754
##
## $mse
## [1] 0.188822
##
## $rmse
## [1] 0.4345366
##
## $rsquare
## [1] 0.9780545
##
## $stdb
##                     x
## 0.28603693 0.01144849
##
## $covb
##                              x
##    0.081817126 -0.0028237861
## x -0.002823786  0.0001310679
```

```
##
## $t
##        [,1]
##    4.308374
## x 17.662722
##
## $prob1
##           [,1]
##   1.765068e-03
## x 2.298018e-07
##
## $prob2
##           [,1]
##   3.530135e-03
## x 4.596036e-07
##
## $H
##              [,1]       [,2]        [,3]        [,4]        [,5]       [,6]
##  [1,] 0.13563935 0.08942540  0.04403669  0.04981343  0.05352705 0.12326061
##  [2,] 0.08942540 0.13028372  0.17041242  0.16530513  0.16202188 0.10036959
##  [3,] 0.04403669 0.17041242  0.29453145  0.27873448  0.26857929 0.07788733
##  [4,] 0.04981343 0.16530513  0.27873448  0.26429802  0.25501744 0.08074871
##  [5,] 0.05352705 0.16202188  0.26857929  0.25501744  0.24629910 0.08258817
##  [6,] 0.12326061 0.10036959  0.07788733  0.08074871  0.08258817 0.11712909
##  [7,] 0.13687723 0.08833098  0.04065162  0.04671990  0.05062094 0.12387377
##  [8,] 0.16741145 0.06133530 -0.04284663 -0.02958711 -0.02106314 0.13899819
##  [9,] 0.20000879 0.03251559 -0.13198666 -0.11105001 -0.09759073 0.15514454
##              [,7]        [,8]        [,9]
##  [1,] 0.13687723  0.16741145  0.20000879
##  [2,] 0.08833098  0.06133530  0.03251559
##  [3,] 0.04065162 -0.04284663 -0.13198666
##  [4,] 0.04671990 -0.02958711 -0.11105001
##  [5,] 0.05062094 -0.02106314 -0.09759073
##  [6,] 0.12387377  0.13899819  0.15514454
##  [7,] 0.13817757  0.17025277  0.20449522
##  [8,] 0.17025277  0.24033880  0.31516037
##  [9,] 0.20449522  0.31516037  0.43330289
##
## $vresid
##               [,1]         [,2]         [,3]         [,4]         [,5]
##  [1,]  0.163210321 -0.016885483 -0.008315096 -0.009405873 -0.010107086
##  [2,] -0.016885483  0.164221583 -0.032177618 -0.031213249 -0.030593297
##  [3,] -0.008315096 -0.032177618  0.133207994 -0.052631207 -0.050713683
##  [4,] -0.009405873 -0.031213249 -0.052631207  0.138916731 -0.048152907
##  [5,] -0.010107086 -0.030593297 -0.050713683 -0.048152907  0.142315323
##  [6,] -0.023274318 -0.018951988 -0.014706843 -0.015247134 -0.015594464
##  [7,] -0.025845434 -0.016678833 -0.007675921 -0.008821747 -0.009558348
##  [8,] -0.031610967 -0.011581455  0.008090387  0.005586698  0.003977184
##  [9,] -0.037766064 -0.006139659  0.024921987  0.020968687  0.018427279
##              [,6]         [,7]         [,8]         [,9]
##  [1,] -0.02327432 -0.025845434 -0.031610967 -0.037766064
##  [2,] -0.01895199 -0.016678833 -0.011581455 -0.006139659
##  [3,] -0.01470684 -0.007675921  0.008090387  0.024921987
##  [4,] -0.01524713 -0.008821747  0.005586698  0.020968687
```

```
##  [5,] -0.01559446 -0.009558348  0.003977184  0.018427279
##  [6,]  0.16670547 -0.023390094 -0.026245919 -0.029294705
##  [7,] -0.02339009  0.162731049 -0.032147472 -0.038613199
##  [8,] -0.02624592 -0.032147472  0.143440760 -0.059509216
##  [9,] -0.02929471 -0.038613199 -0.059509216  0.107004890
##
## $vpred
##              [,1]        [,2]         [,3]         [,4]         [,5]       [,6]
##  [1,] 0.025611696 0.016885483  0.008315096  0.009405873  0.010107086 0.02327432
##  [2,] 0.016885483 0.024600434  0.032177618  0.031213249  0.030593297 0.01895199
##  [3,] 0.008315096 0.032177618  0.055614023  0.052631207  0.050713683 0.01470684
##  [4,] 0.009405873 0.031213249  0.052631207  0.049905285  0.048152907 0.01524713
##  [5,] 0.010107086 0.030593297  0.050713683  0.048152907  0.046506694 0.01559446
##  [6,] 0.023274318 0.018951988  0.014706843  0.015247134  0.015594464 0.02211655
##  [7,] 0.025845434 0.016678833  0.007675921  0.008821747  0.009558348 0.02339009
##  [8,] 0.031610967 0.011581455 -0.008090387 -0.005586698 -0.003977184 0.02624592
##  [9,] 0.037766064 0.006139659 -0.024921987 -0.020968687 -0.018427279 0.02929471
##              [,7]         [,8]         [,9]
##  [1,] 0.025845434  0.031610967  0.037766064
##  [2,] 0.016678833  0.011581455  0.006139659
##  [3,] 0.007675921 -0.008090387 -0.024921987
##  [4,] 0.008821747 -0.005586698 -0.020968687
##  [5,] 0.009558348 -0.003977184 -0.018427279
##  [6,] 0.023390094  0.026245919  0.029294705
##  [7,] 0.026090968  0.032147472  0.038613199
##  [8,] 0.032147472  0.045381257  0.059509216
##  [9,] 0.038613199  0.059509216  0.081817126
##
## $h
## [1] 0.1356394 0.1302837 0.2945315 0.2642980 0.2462991 0.1171291 0.1381776
## [8] 0.2403388 0.4333029
##
## $lowerrm95
##           [,1]
##  [1,] 4.0084275
##  [2,] 6.2807426
##  [3,] 8.3183089
##  [4,] 8.0646082
##  [5,] 7.9009220
##  [6,] 4.6418300
##  [7,] 3.9442397
##  [8,] 2.3260921
##  [9,] 0.5559843
##
## $uppperrm95
##           [,1]
##  [1,] 4.765280
##  [2,] 7.022503
##  [3,] 9.433590
##  [4,] 9.121098
##  [5,] 8.920804
##  [6,] 5.345147
##  [7,] 4.708141
##  [8,] 3.333558
```

```
##  [9,] 1.908724
##
## $lower95
##              [,1]
##  [1,] 3.291867514
##  [2,] 5.559221477
##  [3,] 7.706867963
##  [4,] 7.437504287
##  [5,] 7.263767347
##  [6,] 3.907462538
##  [7,] 3.229981064
##  [8,] 1.685475793
##  [9,] 0.002205941
##
## $upper95
##             [,1]
##  [1,]  5.481840
##  [2,]  7.744024
##  [3,] 10.045031
##  [4,]  9.748202
##  [5,]  9.557958
##  [6,]  6.079514
##  [7,]  5.422400
##  [8,]  3.974174
##  [9,]  2.462502
```

```
my.regress.plot(result)
```

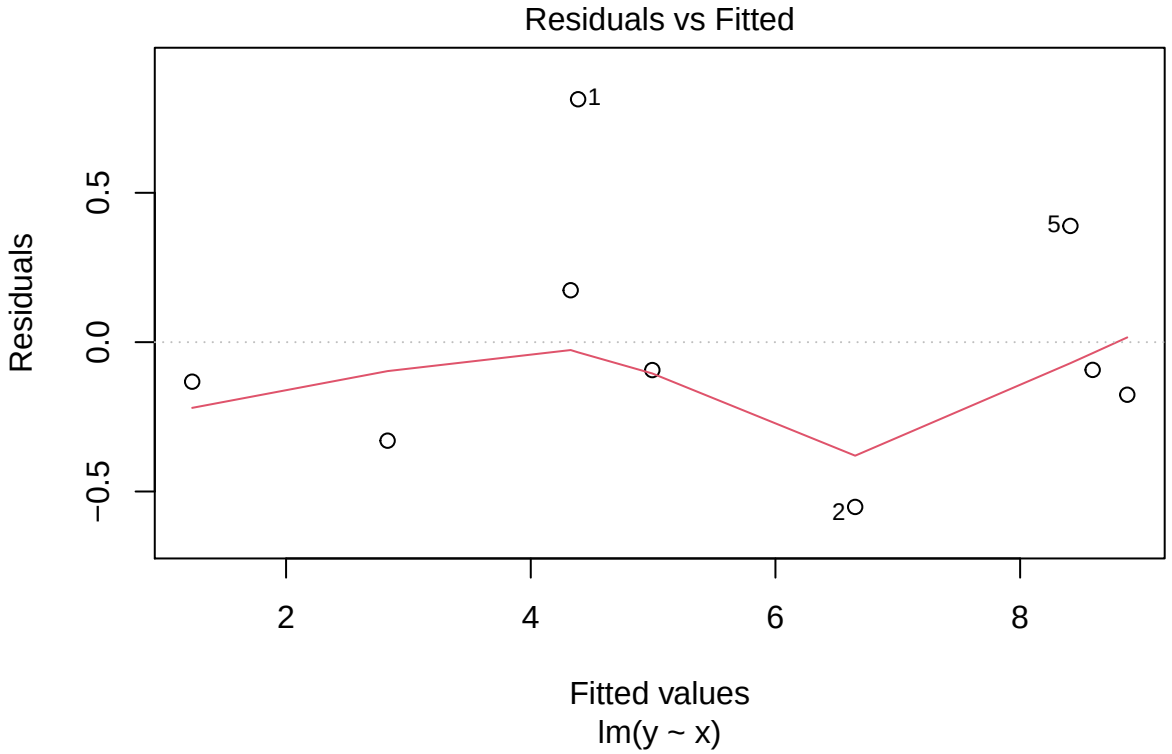


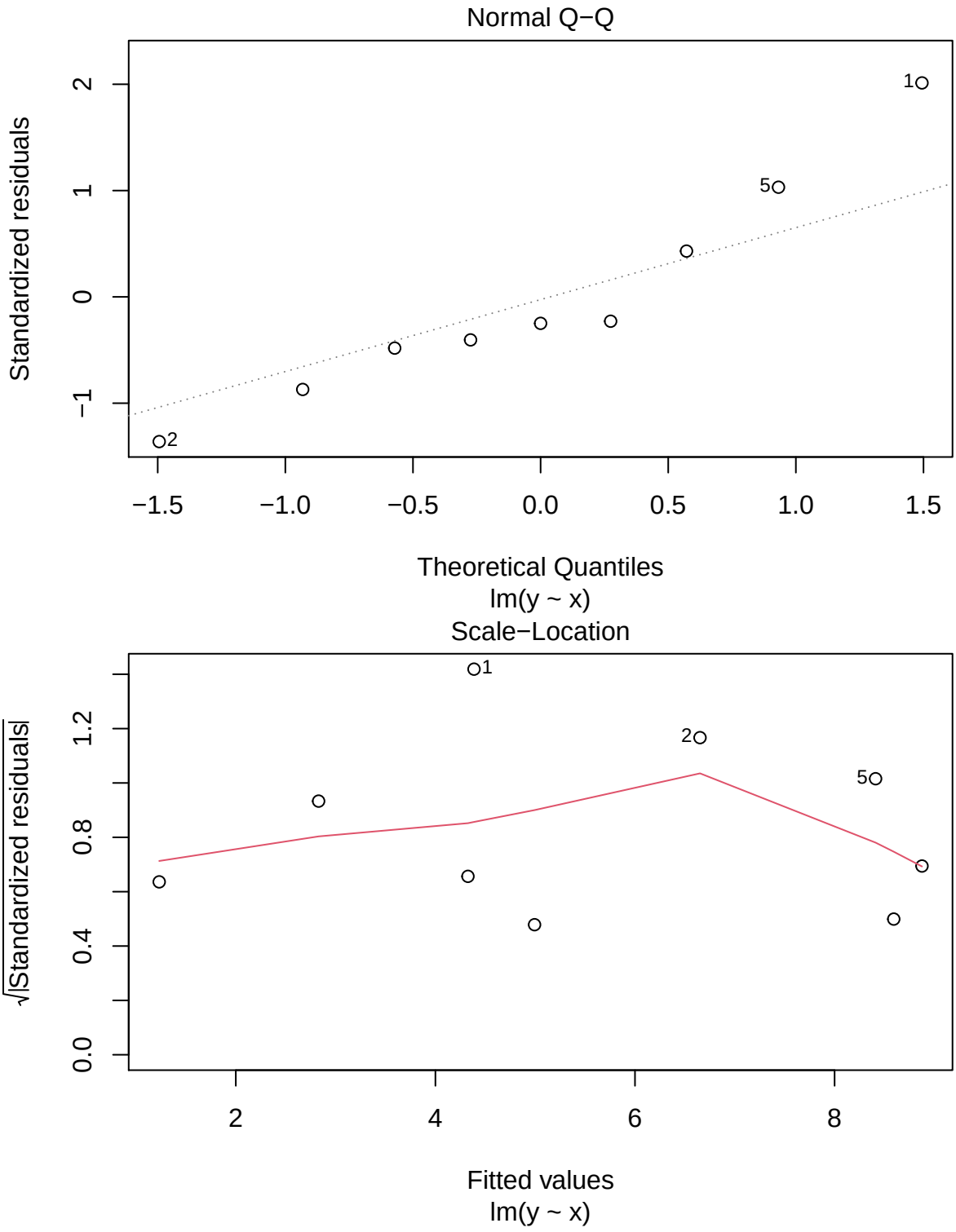
R packages for Regression

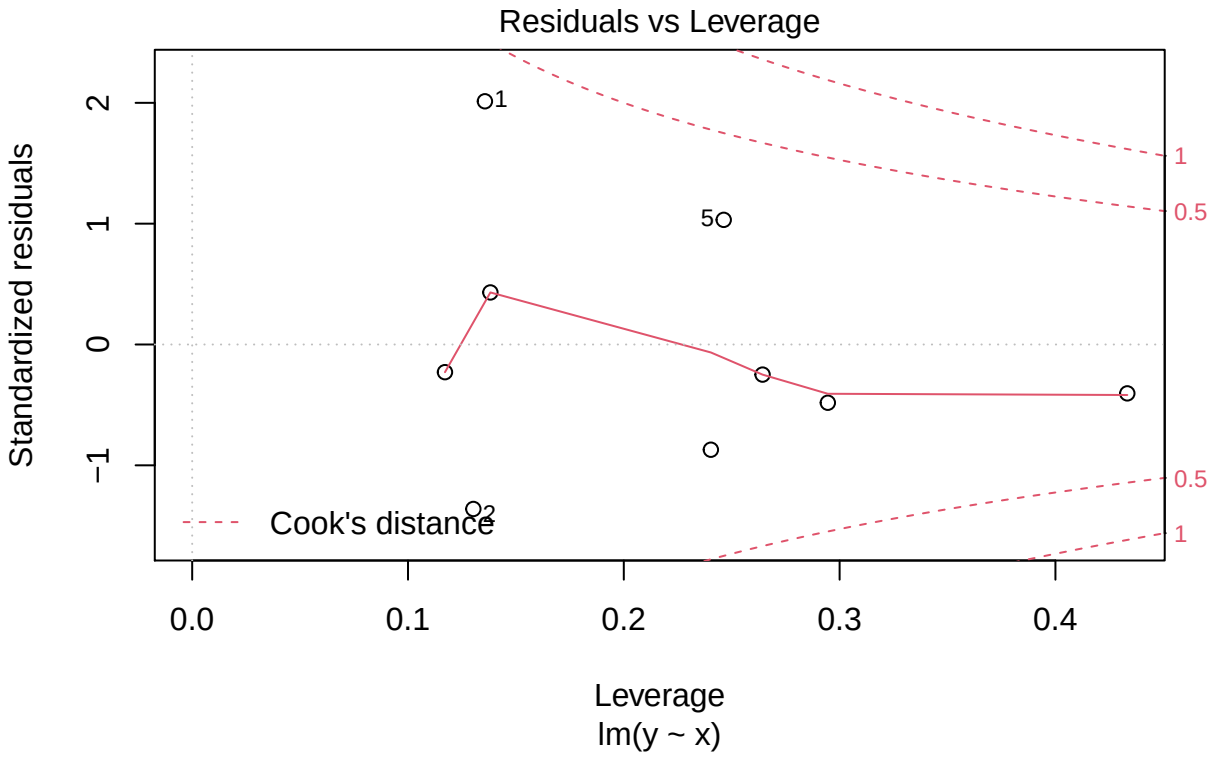
```
result<-lm(y~x)
summary(result)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55162 -0.17595 -0.09349  0.17381  0.81315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.23235    0.28604   4.308  0.00353 **
## x            0.20221    0.01145  17.663  4.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
##
## Residual standard error: 0.4345 on 7 degrees of freedom
## Multiple R-squared:  0.9781, Adjusted R-squared:  0.9749
## F-statistic:   312 on 1 and 7 DF,  p-value: 4.596e-07
```

```
plot(result)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ x)



Assignment

Non matrix Notation

1. Sketch a simple linear regression line where you indicate the observed data, the location of $\hat{\beta}_0$, $\bar{x}$, $\bar{y}$. Indicate what the value $\hat{\beta}_1$ means. Add the confidence intervals for the mean of Y and for an observation y_i at $X = x_i$.
2. Using the expression for $\hat{\sigma}_{\hat{\mu}_{y|x^*}}$, answer the following
   (a) At which value of x is the estimate of the line most precise? What does this mean if the regression line is used to predict the dependent variable y?
   (b) Assume that you can select the locations for x_i, where should you place $\bar{x}$?
   (c) What effect does SS_{xx} have? What does this mean?
   (d) Suppose that you do not have control of where x_i is placed, what implication does this have in terms of the accuracy of the line? Explain.
3. Suppose that $y_i = x_i\beta + e_i$ for $i = 1, 2, \ldots, n$. Find
   (a) The least squares estimate for β, given by $\hat{\beta}$.
   (b) $E(\hat{\beta})$
   (c) $Var(\hat{\beta})$.
   (d) Assume that e_i are i.i.d $N(0, \sigma^2)$. Find the distribution for $\hat{\beta}$.

(e) Answer the above questions when $x_i = c \neq 0$. Suppose $c = 1$, what does this imply about $\hat{\beta}_1$ or the need for x_i in the usual least squares line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$?

4. Reproduce another example using the R-code given in this document. There are many data sets, look in elementary statistics texts.

Matrix Notation

1. Under what conditions would the rank of X be less than 2? What does this mean? Can this be avoided, assuming that the choice of the x_i is yours?
2. Show that H and $(I - H)$ are idempotent matrices. Furthermore, H and $I - H$ are orthogonal since $H(I - H) = (I - H)H = 0$.
3. Find HX and $X'H$, $(I - H)X$, $X'(I - H)$, $H\mathbf{j_n}$. If $H\mathbf{j_n} = \mathbf{j_n}$, what does this imply about the rows (or columns) of H?
4. If $\hat{y} = Hy$, what can you say about $\hat{y}_i$ in terms of y_i? What is the value of $\sum_{i=1}^{n} h_{ii}$ when $H = (h_{ij})$. What is the value of $\sum_{j=1}^{n} h_{ij}$?
5. Suppose h_{ii} is large, what does this mean? What does large mean? Explain your answer.

SAS Program

Code

```
title 'Heat Data Example';
data heat; set sasuser.heat;
proc print;
run;

proc reg data=heat plots=fitplot;
model y = x;
run;
```

Output

Obs	x	y
1	15.60	5.20
2	26.80	6.10
3	37.80	8.70
4	36.40	8.50
5	35.50	8.80
6	18.60	4.90
7	15.30	4.50
8	7.90	2.50
9	0.00	1.10

Heat Data Example

The REG Procedure

Model: MODEL1

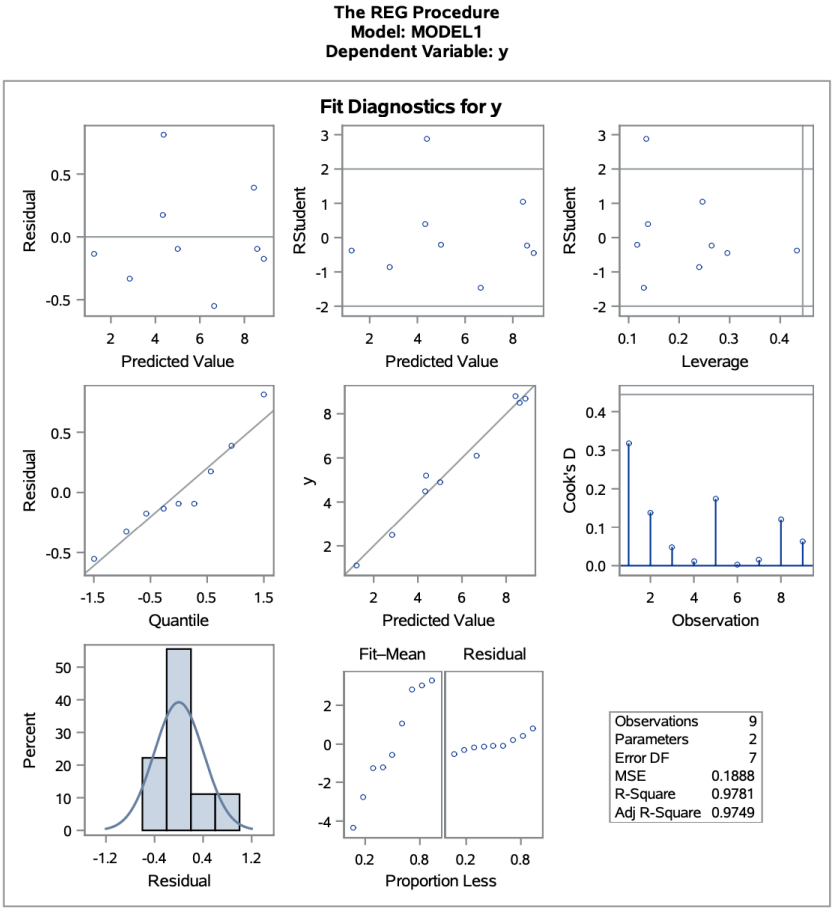
Dependent Variable: y

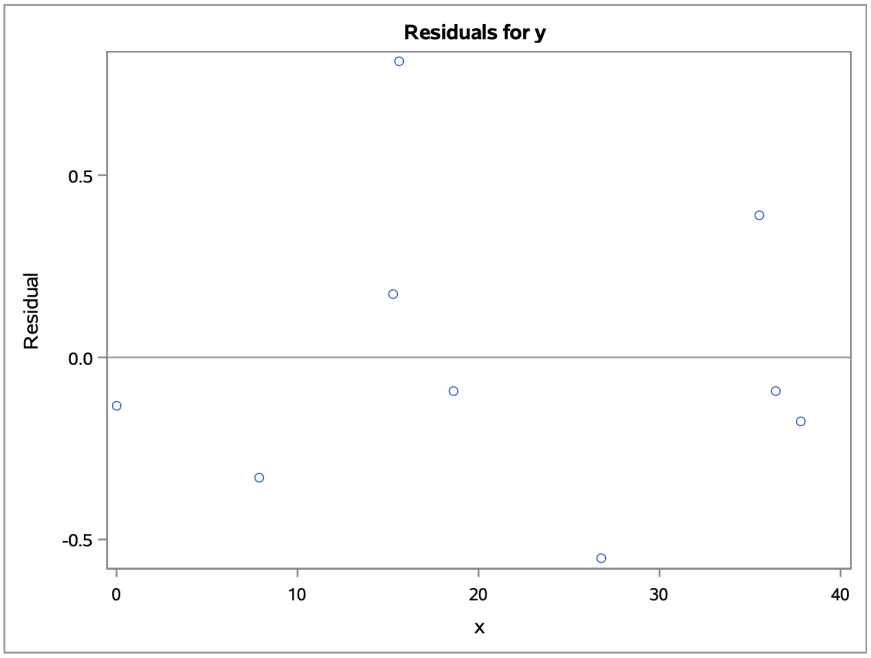
Number of Observations Read	9
Number of Observations Used	9

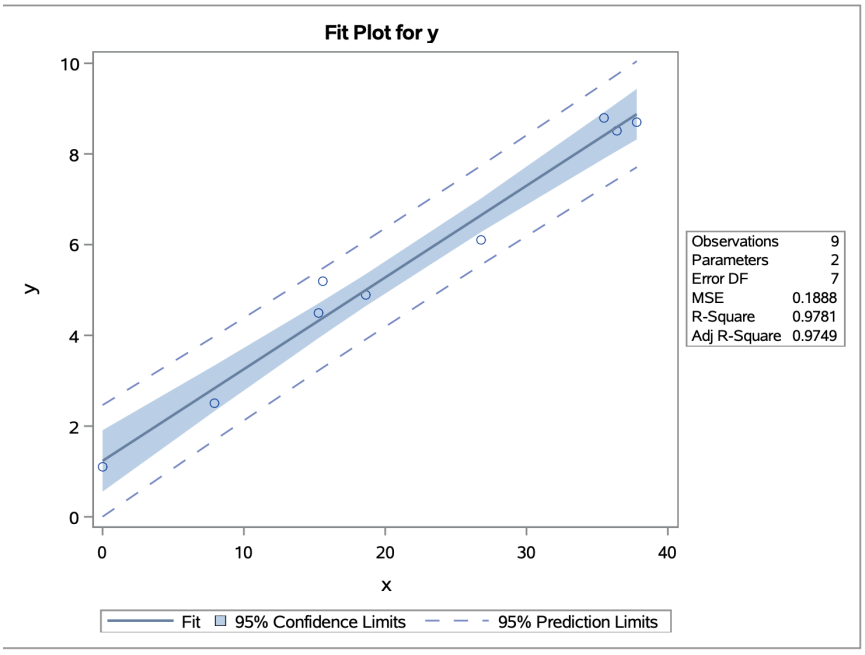
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	58.90713	58.90713	311.97	<.0001
Error	7	1.32175	0.18882		
Corrected Total	8	60.22889			

Root MSE	0.43454	R-Square	0.9781
Dependent Mean	5.58889	Adj R-Sq	0.9749
Coeff Var	7.77501		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.23235	0.28604	4.31	0.0035
x	1	0.20221	0.01145	17.66	<.0001




Residuals for y
Residual
0.5
0.0
-0.5
0
10
20
30
40
x


Fit Plot for y
y
10
8
6
4
2
0
0
10
20
30
40
x
Observations 9
Parameters 2
Error DF 7
MSE 0.1888
R-Square 0.9781
Adj R-Square 0.9749
Fit
95% Confidence Limits
95% Prediction Limits