Linear Regression Example

jdt

1/7/2021

Contents

Theory **1**

R **7**

SAS **11**

Theory

In this document, we consider a modeling problem whereby we are interested in determining the relationship between two (or more) random variables on a single population of interest. These models are called *regression models*. The simplest of these models is the simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

where the problem of interest is to determine the slope, β_1, and y-intercept, β_0, that best fits the observed data in the least squared error sense.

Least Squares Solution – Simple Linear Model

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 least squares problem consists of finding the linear equation given by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

which minimizes

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

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

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

Equation(2) is called **the normal equations** for the linear least squares problem. From which, the unique solution is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3}$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{4}$$

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[1]

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

[1] Note if the estimated slope $\hat{\beta}_1$ is zero then the predicted value for every y_i is $\bar{y}$.

Inference in Linear Regression

The procedure for finding the least squares estimates for β_0 and β_1 is a mathematical problem, in that, it provides the solution to an optimization problem. In order to make the problem a statistical problem it is necessary to make a distributional assumption concerning the dependent variable y [in this chapter]. The assumptions used in linear regression 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

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

- 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 $df = n - 2$.

Problems

1. Using the expression for $\hat{\sigma}_{\hat{\mu}_{y|x^*}}$ [in red above], 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.
2. 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 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$?

Regression Example with Excel

Given 9 pairs of observations where the independent variable X = heating degree days and the dependent variable Y = gas consumption for house.

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	$(y-\hat{y})$	$\hat{y}$
15.6	5.2	-5.94	-0.39	35.34	0.151	2.312	0.813	4.387
26.8	6.1	5.256	0.511	27.62	0.261	2.686	-0.55	6.652
37.8	8.7	16.26	3.111	264.2	9.679	50.57	-0.18	8.876
36.4	8.5	14.86	2.911	220.7	8.475	43.25	-0.09	8.593
35.5	8.8	13.96	3.211	194.8	10.31	44.81	0.389	8.411
18.6	4.9	-2.94	-0.69	8.67	0.475	2.028	-0.09	4.993
15.3	4.5	-6.24	-1.09	38.99	1.186	6.8	0.174	4.326
7.9	2.5	-13.6	-3.09	186.2	9.541	42.15	-0.33	2.83
0	1.1	-21.5	-4.49	464.2	20.15	96.71	-0.13	1.232
193.9	50.3	0.002	0	1441	60.23	291.3	0	50.3
21.54	5.589	0	0	160.1	6.692	32.37	0	5.589

The last two lines of the above table denote the column totals and averages. Hence $\bar{x}$ = 21.54 and $\bar{y}$ = 5.589. The sample variances are computed using the totals under $(x-\bar{x})^2$ = 1441 and $(y-\bar{y})^2$ = 60.23. That is SS_{xx} = 1441 and SS_{yy} = 60.23 and the sample variance of x is given by SS_{xx}/(n-1) = 1441/8 = 180.08 and the sample variance for y is given by SS_{yy}/8 = 60.23/8 = 7.529. In order to find the least square estimates for the y intercept and for the slope of the line one computes the **estimate for the slope**

$$\hat{\beta}_1 = SS_{xy}/SS_{xx} = 291.3/1441 = .20221.$$

where SS_{xy} is given as the total under the $(x - \bar{x})(y - \bar{y})$ column. The **estimate for the y intercept**

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 5.589 - .20221 * 21.54 = 1.233.$$

The residuals are calculated by computing $y - \hat{y}$ where $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The predicted values are in the $\hat{y}$ column and the residuals are in the $(y - \hat{y})$ column. The residual sum of squares is found by summing the values under $(y - \hat{y})^2$. One should get SSE = 1.323. From here one estimates the variance of y by computing SSE/(n-2) = 1.323/7 = .189. The standard deviation of y is given by $\sqrt{.189} = .435$. (this number is often called the root mean square error, denoted by $s = \hat{\sigma}$).

Suppose that one wish to find a 95% CI for the slope of the line. It is given by $\hat{\beta}_1 \pm (t_{.025}(n-2)) * \hat{\sigma}/\sqrt{SS_{xx}} = .20221 \pm (2.365)(.01146) = (.17511, .22931)$. where .01146 = .435/$\sqrt{1441}$ and 2.365 is obtained from the t distribution table under the .025 column with 7 degrees of freedom.

Suppose that you want to test the hypothesis that the slope of the line is equal to zero at the .05 level. One can reject this hypothesis since 0 is not contained in the above interval. Note: one can only use confidence intervals to formulate test of hypothesis when the level of the CI is comparable to the specified type I error.

Suppose that one wanted a 95% CI for the line at $x = 20$. The interval is given by

$$\begin{aligned}
&= (\hat{\beta}_0 + \hat{\beta}_1 \cdot 20) \pm (t_{.025}(n-2))\hat{\sigma}\sqrt{1/n + (20 - \bar{x})^2/SS_{xx}} \\
&= 1.233 + (.20221)(20) \pm 2.365(.435)\sqrt{1/9 + (20 - 21.54)^2/1440} \\
&= 5.277 \pm (2.365)(.146) \\
&= (4.932, 5.622).
\end{aligned}$$

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	$\sum(\hat{y}_i-\bar{y})^2$	$MS_M = SS_M/1$
Residual	n-2	$\sum(y_i-\hat{y}_i)^2$	$MS_E = SS_E/(n-2)$
Corrected Total	n-1	$\sum(y_i-\bar{y})^2$	
due to β_0	1	$n\bar{y}^2$	
Total	n	$\sum y_i^2$	

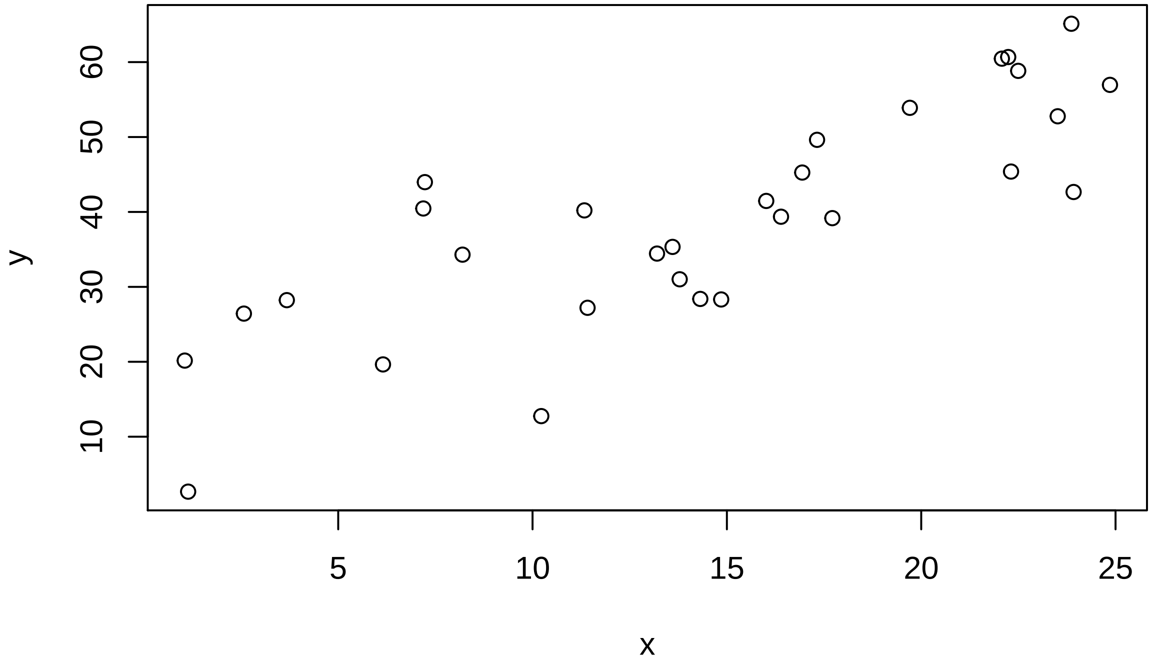
R

Generate the Simulated Data

```
set.seed(123)
n=30          #sample size
beta_0=10     #true y-intercept
beta_1=2     #true slope
sigma= 9      #true sigma

x=25*runif(n)
y=beta_0 + beta_1*x + sigma*rnorm(n)
```

Graph of data

```
plot(y~x)
```



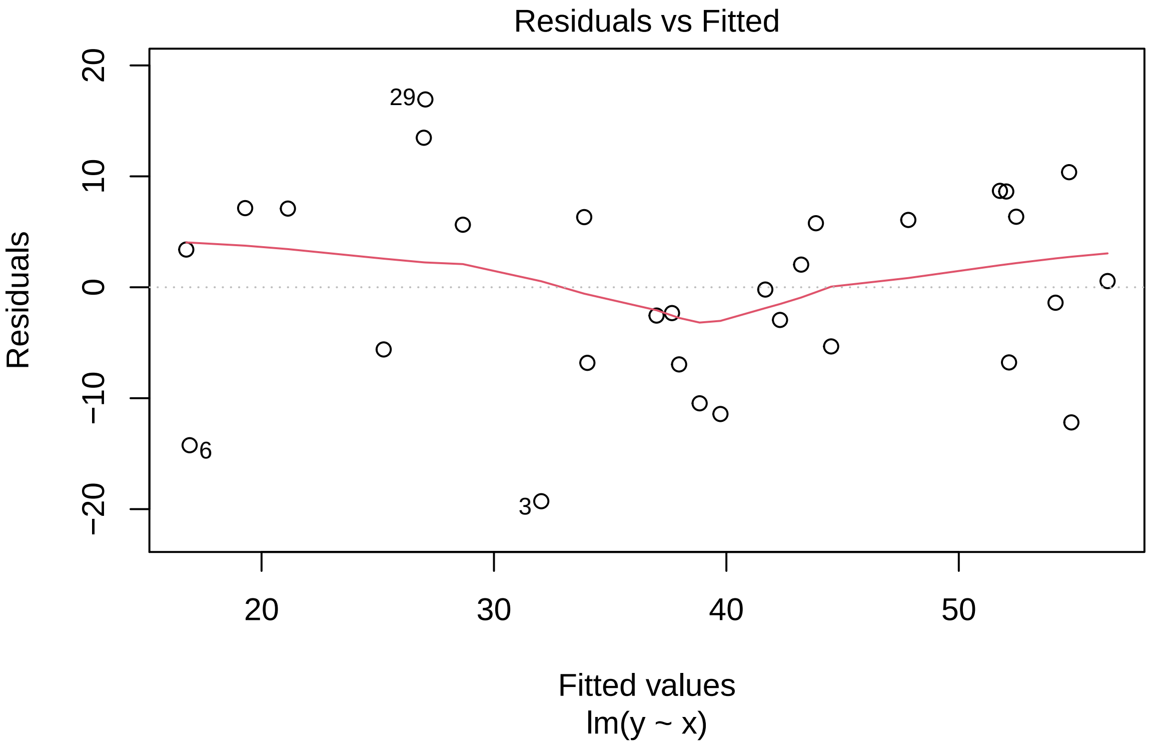
##Results

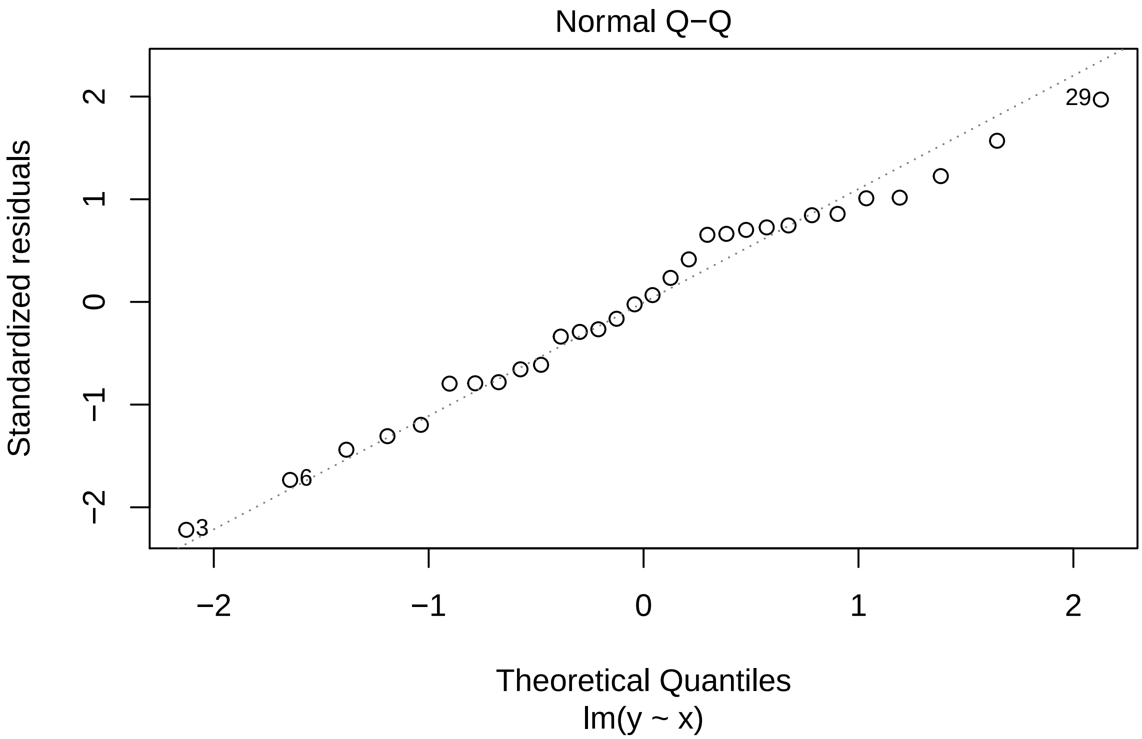
```
result<-lm(y~x)
summary(result)
```

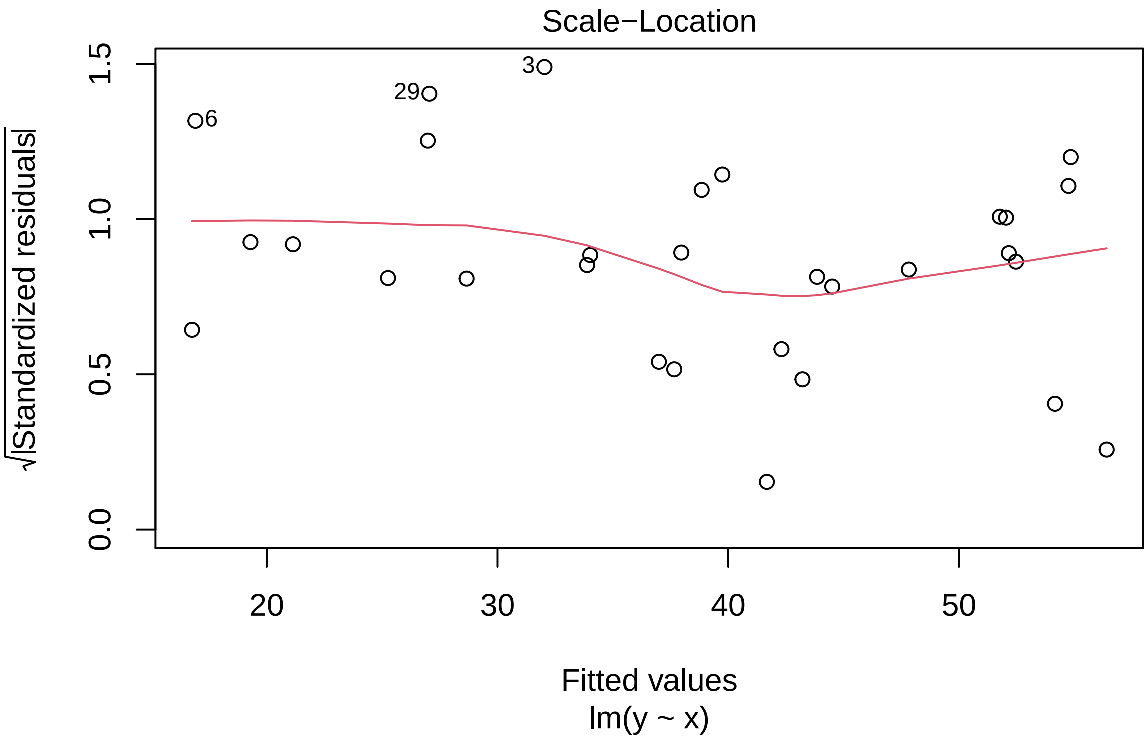
```
##
## Call:
## lm(formula = y ~ x)
##
```

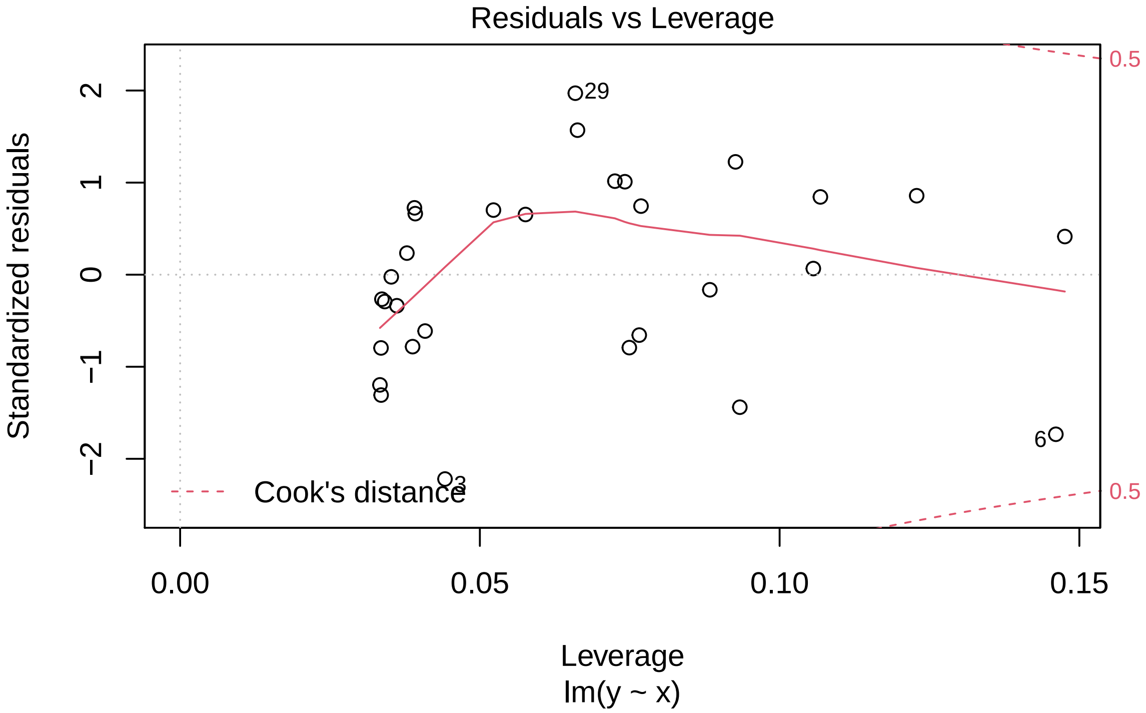
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2864  -6.4826   0.1758   6.3506  16.9309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.0097     3.6258   4.140 0.000289 ***
## x             1.6652     0.2266   7.349  5.3e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.888 on 28 degrees of freedom
## Multiple R-squared:  0.6586, Adjusted R-squared:  0.6464
## F-statistic: 54.01 on 1 and 28 DF,  p-value: 5.296e-08
```

```
plot(result)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
2
1
0
−1
−2
−2
−1
0
1
2
29
6
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
30
40
50
Fitted values
lm(y ~ x)
6
29
3


Residuals vs Leverage
Standardized residuals
−2
−1
0
1
2
0.00
0.05
0.10
0.15
Leverage
lm(y ~ x)
29
6
3
Cook's distance
0.5
0.5

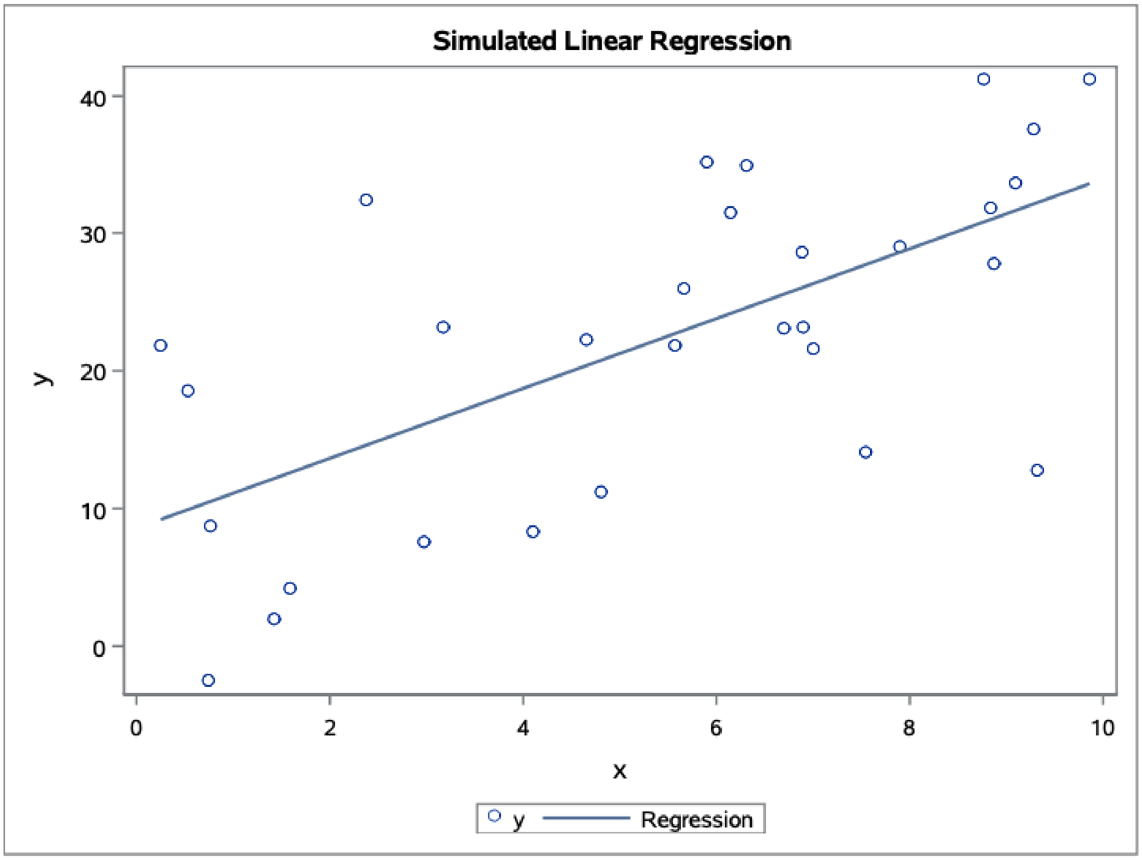
SAS

Code

```
*options center nodate pagesize=100 ls=70;
libname LDATA '/home/jacktubbs/my_shared_file_links/jacktubbs/LaTeX/';

/* Simplified LaTeX output that uses plain LaTeX tables */
ods tagsets.simplelatex
file="/home/jacktubbs/my_shared_file_links/jacktubbs/stat5381_ex.tex"
stylesheet="/home/jacktubbs/my_shared_file_links/jacktubbs/sas.sty"(url="sas");

/*
The above will create a new file that can be inputed into
LaTeX (simple.tex) and the new style needed by LaTex (sas.sty).
In this case I asked that these files be placed in My SAS
Files folder in My Documents. You can put these anywhere.
The following example can be found at

http://support.sas.com/rnd/base/ods/odsmarkup/latex.html

*/

title1 'Simulated Linear Regression';

/****************************************************************
 Simple Linear Regression Models
 ****************************************************************/

%let N = 30;                          /* size of each sample      */
%let beta_0 = 10;                     /* true y-intercept         */
%let beta_1 = 2;                      /* true slope               */
%let sigma=9;                         /* true sigma               */
data Reg1(keep=x y);
call streaminit(1);
do i = 1 to &N;
   x = 10*rand("Uniform");            /* explanatory variable     */
   eps = rand("Normal", 0, &sigma);   /* error term N(0,sigma)    */
   y = &beta_0 + &beta_1*x + eps;
   output;
end;
run;

data reg_out; set Reg1;

proc sgplot data=reg_out;
scatter y=y x=x;
```

```
reg y=y x=x;
run;

proc reg data=Reg1 plots=FITPLOT;
   model y = x;

   run;
quit;
```

Output



Simulated Linear Regression

The REG Procedure

Model: MODEL1

Dependent Variable: y

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1715.99788	1715.99788	20.47	0.0001
Error	28	2346.79904	83.81425		
Corrected Total	29	4062.79692			

Root MSE	9.15501	R-Square	0.4224
Dependent Mean	22.42470	Adj R-Sq	0.4017
Coeff Var	40.82557		

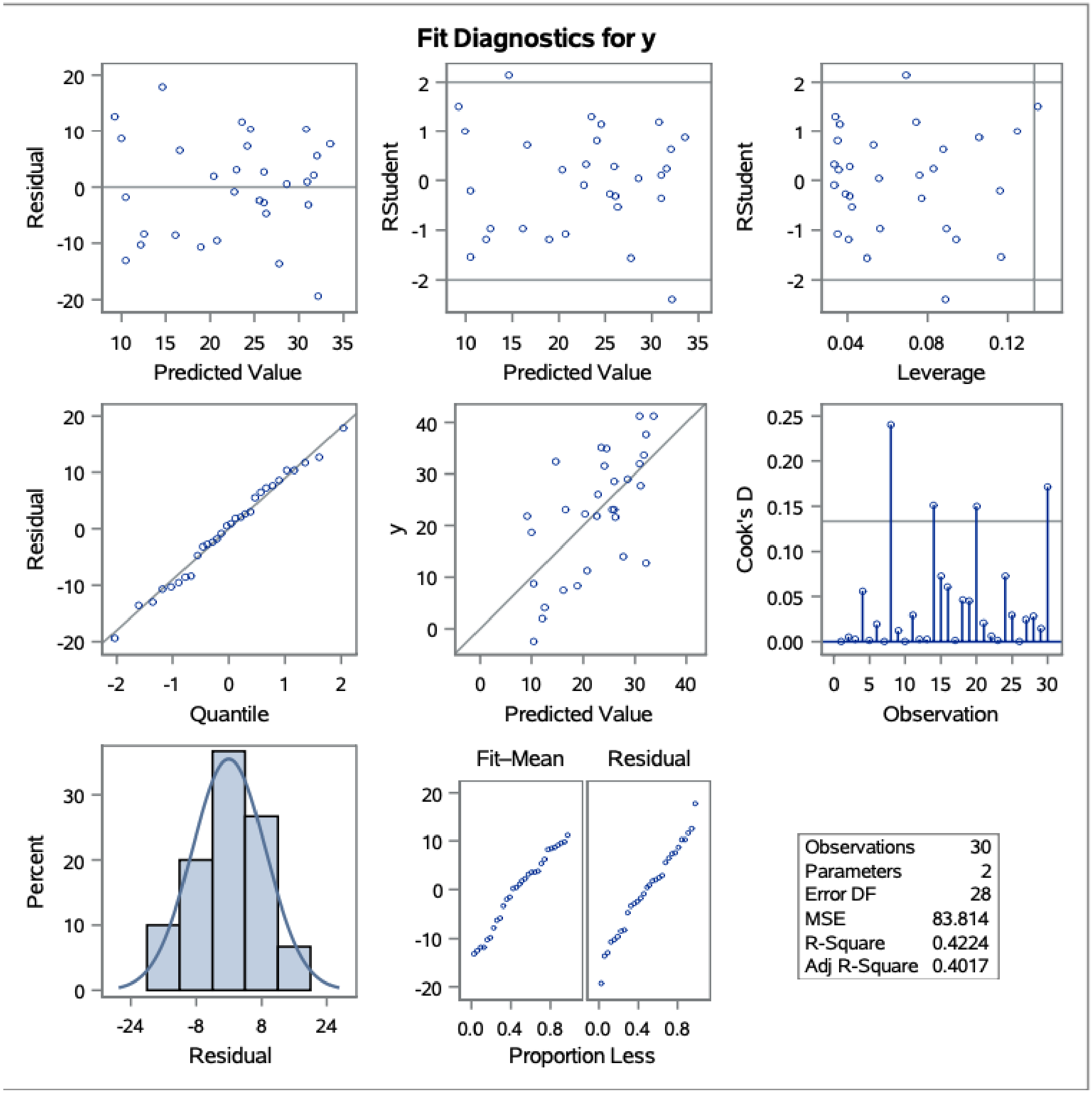
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.56882	3.48869	2.46	0.0205
x	1	2.53618	0.56051	4.52	0.0001

Simulated Linear Regression

The REG Procedure

Model: MODEL1

Dependent Variable: y



Observations	30
Parameters	2
Error DF	28
MSE	83.814
R-Square	0.4224
Adj R-Square	0.4017