Linear Regression

Linear regression

- Linear regression models the relationship between a variable *y* and one (or more) explanatory variable *x*

- It is an important tool for scalability studies: what is the performance of a system with increasing input size?

- Examples:
  - Response time of a database with increasing number of queries
  - Response time of a webserver with increasing number of requests
  - CPU-time of an algorithm to solve an optimization problem with increasing instance size

Linear regression

- A linear regression model has the following form:

$$y = a + bx$$

where x is the input variable, y is the predicted output response, and a and b are regression parameters that we wish to estimate from our set of measurements.

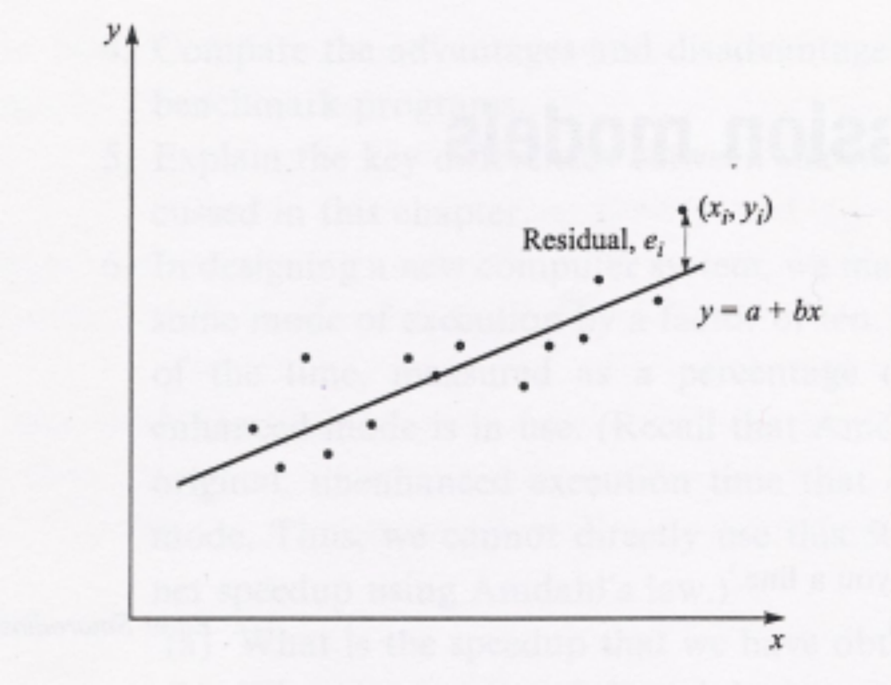
In other words: we would like to fit a line to a set of points, where a is the y-intercept and b is the slope.

Linear regression

- If y_i is the value measured when we set the input value to x_i, each pair (x_i, y_i), the expression can be written as follows:

$$y_i = a + bx_i + e_i$$

where e_i (residual) is the difference between the measured value of y_i and the value that would have been predicted for y_i .



Linear regression

- In order to find the regression parameters *a* and *b*, we minimize the sum of squares of the residuals (SSE). That is, we wish to find *a* and *b* that minimizes

$$\min \text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

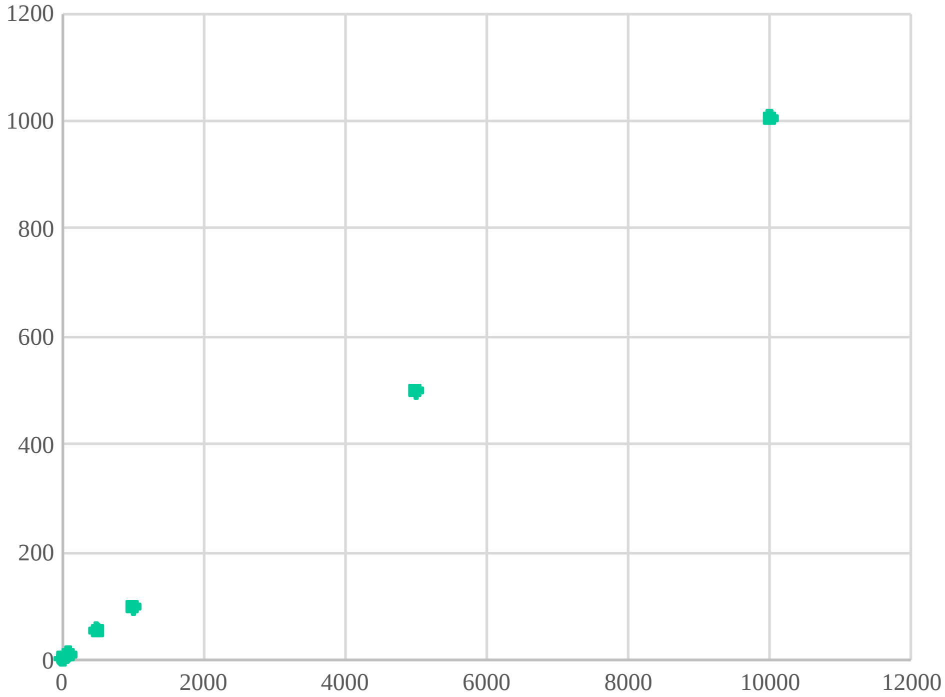
By calculus, *a* and *b* can be estimated as follows:

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

$$\hat{a} = \bar{y} \text{ - } \hat{b}\bar{x}$$

Example of linear regression

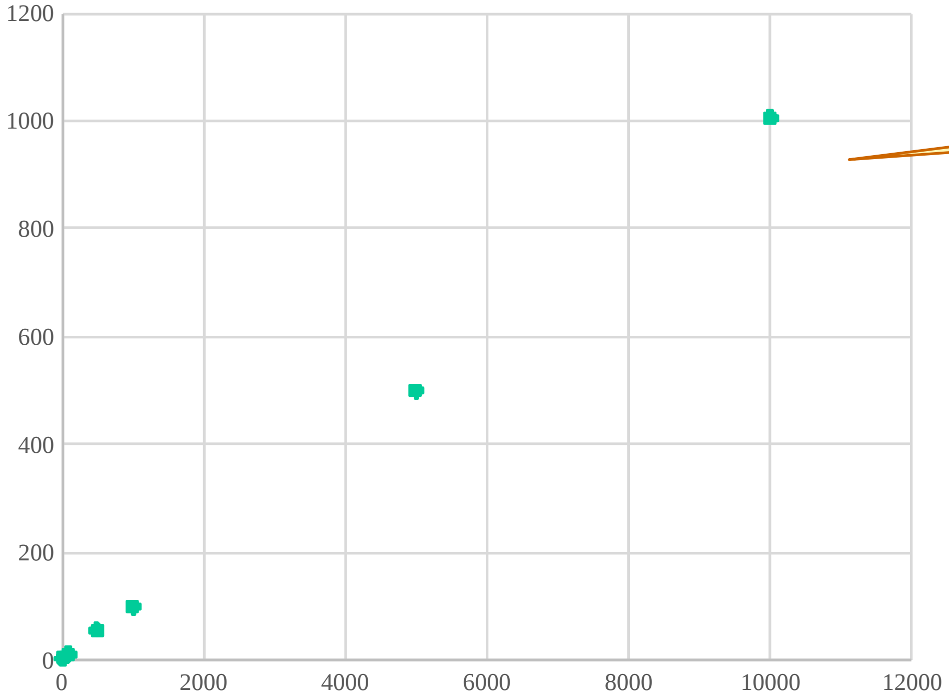
Develop a regression model to relate the time required to perform a file-read operation to the number of bytes read.



File size in bytes (x_i)	Time in µs (y_i)
10	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1

Example of linear regression

Develop a regression model to relate the time required to perform a file-read operation to the number of bytes read.



It suggest a linear relationship

	(y_i)
10	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1

Example of linear regression

Develop a regression model to relate the time required to perform a file-read operation to the number of bytes read.

File size in bytes (x_i)	Time in µs (y_i)
10	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1

$b = 0.1002$

$a = 2.24$

$y = 2.24 + 0.1002\, x$

The time required to read a file is appr. 2.24 µs + 0.1002 µs per byte read.

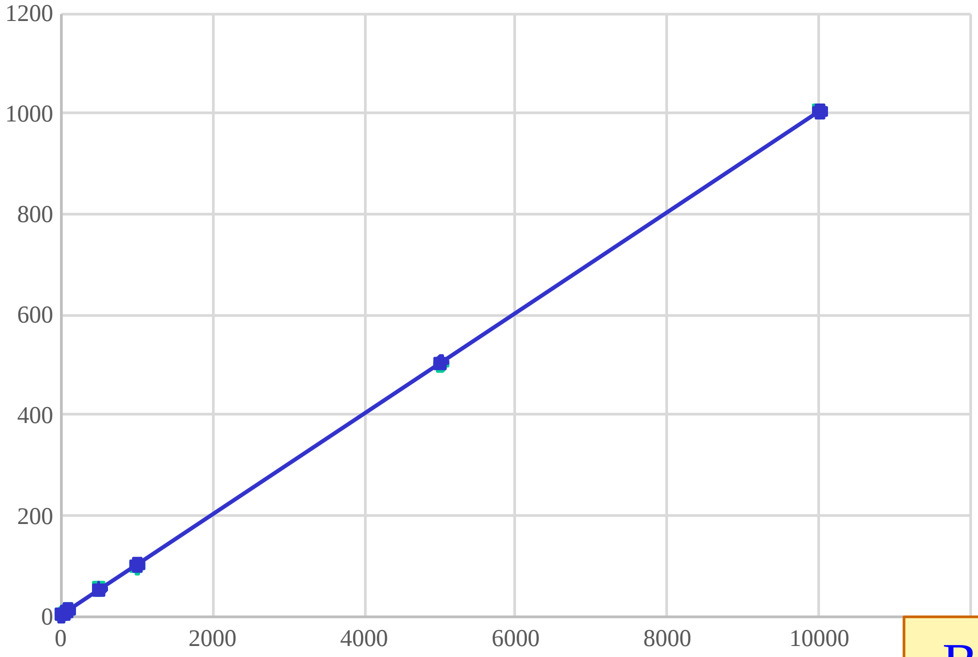
Example of linear regression

In R

```
> D = read.table("regr.in",header=TRUE)
> lr.out = lm(D$time~D$size)
> summary(lr.out)

Call:
lm(formula = D$time ~ D$size)

Residuals:
      1       2       3       4       5       6       7
 0.5584  0.8497 -0.3612  3.2518 -2.8570 -3.1270  1.6854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.239467   1.163822   1.924    0.112
D$size      0.100218   0.000274 365.717  2.9e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.55 on 5 degrees of freedom
Multiple R-squared:     1,    Adjusted R-squared:     1
F-statistic: 1.337e+05 on 1 and 5 DF,  p-value: 2.901e-12
```

Example of linear regression

Develop a regression model to relate the time required to perform a file-read operation to the number of bytes read.



File size in bytes (x_i)	Time in µs (y_i)
10	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1

$$y = 2.24 + 0.1002\, x$$

But how good is the fit?

Validation of the model

- We are interested in knowing how good is the model.
- The total variation of in the measured system outputs (SST) is partitioned into two components:
  1. SSR: portion of SST that is explained by the regression model
  2. SSE: portion of SST that is due to measurement error

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

$$\text{SSR} = \text{SST} - \text{SSE}$$

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

where $\hat{y}_i$ is the estimate of y_i according to the model.

Correlation

- We are interested in knowing how good is the model.
- The total variation of in the measured system outputs (SST) is partitioned into two components:
  1. SSR: portion of SST that is explained by the regression model
  2. SSE: portion of SST that is due to measurement error

The coefficient of determination r^2 gives the fraction of the total variation explained by the regression model

$$r^2 = \frac{SSR}{SST}$$

Correlation

- We are interested in knowing how good is the model.
- The total variation of in the measured
partitioned into two components:
  1. SSR: portion of SST that is expla
model
  2. SSE: portion of SST that is due to

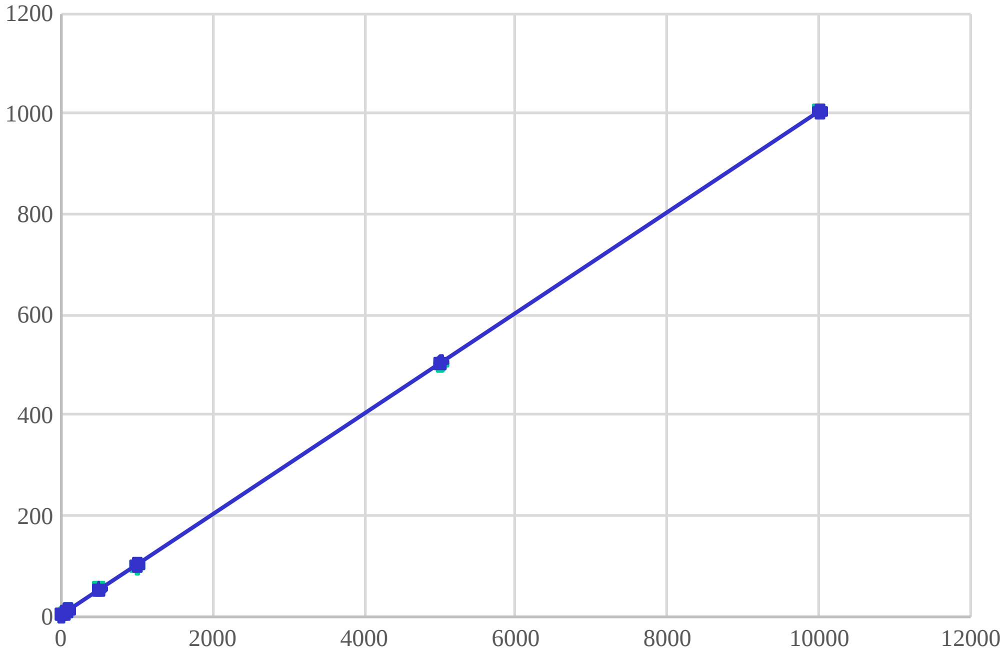
The coefficient of determination r^2 giv e fraction of the total variation explained by the regres on model

$$r^2 = \frac{SSR}{SST}$$

Note that $0 \leq r^2 \leq 1$. If there is a perfect relationship between input and output, then the total variation is explained by the model, implying $r^2=1$.

Example of linear regression

Coefficient of determination: $r^2 = 0.9996$



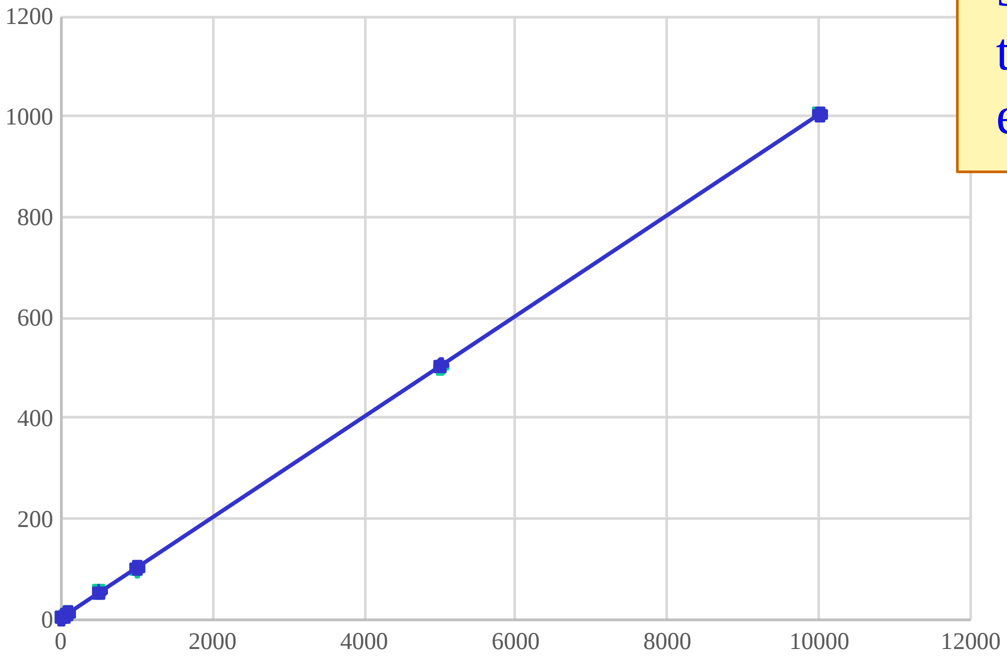
File size in bytes (x_i)	Time in µs (y_i)
10	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1

$$y = 2.24 + 0.1002\, x$$

Example of linear regression

Coefficient of determination: $r^2 = 0.9996$

99.96% of the variation in the time required to read a file is explained by this model



	e in μs (y_i)
	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1

$$y = 2.24 + 0.1002\,x$$

Example of linear regression

In R

```
> D = read.table("regr.in",header=TRUE)
> lr.out = lm(D$time~D$size)
> summary(lr.out)

Call:
lm(formula = D$time ~ D$size)

Residuals:
      1       2       3       4       5       6       7
 0.5584  0.8497 -0.3612  3.2518 -2.8570 -3.1270  1.6854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.239467   1.163822   1.924    0.112
D$size      0.100218   0.000274 365.717  2.9e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.55 on 5 degrees of freedom
Multiple R-squared:     1,     Adjusted R-squared:     1
F-statistic: 1.337e+05 on 1 and 5 DF,  p-value: 2.901e-12
```

Transformations

- The relationship you are trying to model may not be linear. Although you could still apply a linear model, it would give wrong predictions.
- In many cases, it is possible to *transform* the nonlinear data into a linear form. For instance, if you expect an exponential behavior of your system:

$$y = ab^x$$

By taking the logarithm of both sides

$$\ln y = \ln a + (\ln b)\, x$$

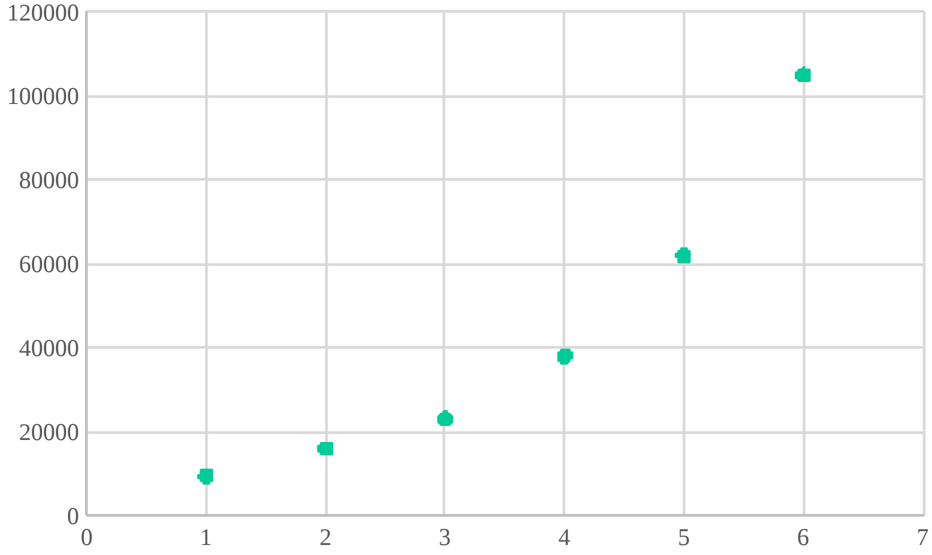
the expression has a linear form.

$$y' = a' + b' x$$

See Chapter 8 “Transformations” in J. Faraway, *Practical Regression and ANOVA in R*

Example of a nonlinear model

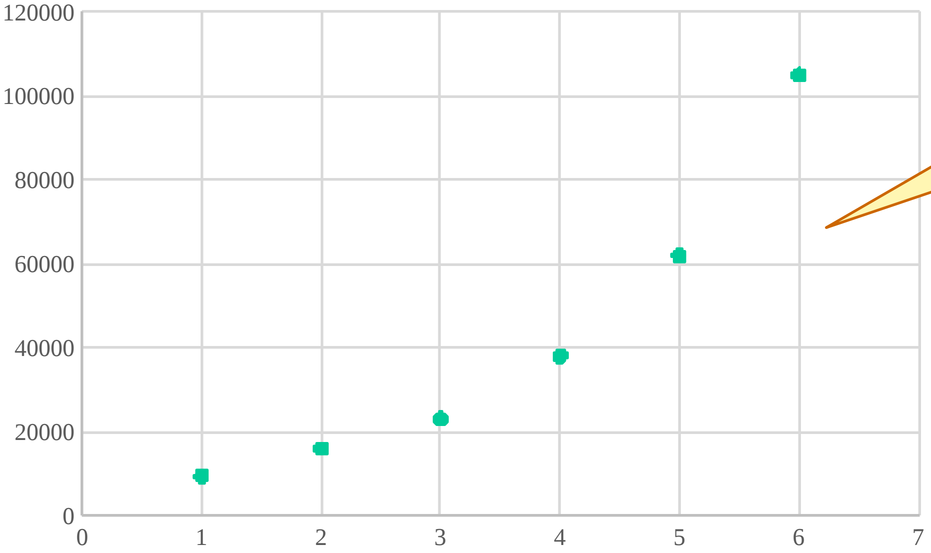
Example 1: Estimated number of transistors in the following 6 years



Year (x_i)	Number of transistors (y_i)
1	9500
2	16000
3	23000
4	38000
5	62000
6	105000

Example of a nonlinear model

Example 1: Estimated number of transistors in the following 6 years



This is not a linear relationship. Since it almost duplicates every year, it may be an exponential relationship

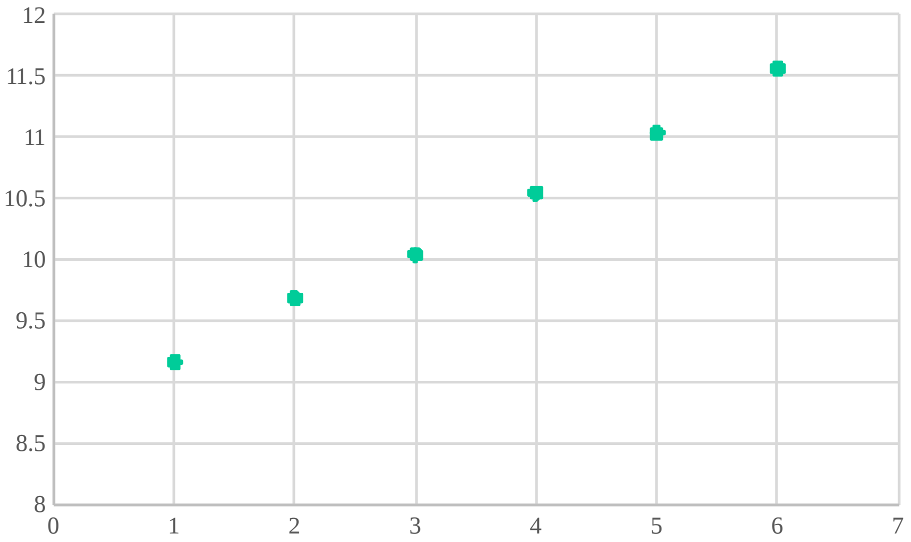
	ors (y_i)
1	9500
2	16000
3	23000
4	38000
5	62000
6	105000

$$y = ab^x$$

Example of a nonlinear model

Example 1: Estimatec following 6 years

Logarithmic transformation of the response variable



Year (x_i)	Transformed data (y_i'=ln y_i)
1	9.1590
2	9.6803
3	10.0432
4	10.5453
5	11.0349
6	11.5617

$$y' = a' + b'x$$

Example of a nonlinear model

Example 1: Estimated number of transistors in the following 6 years

$b' = 0.474$ $b = e^{b'} = 1.61$

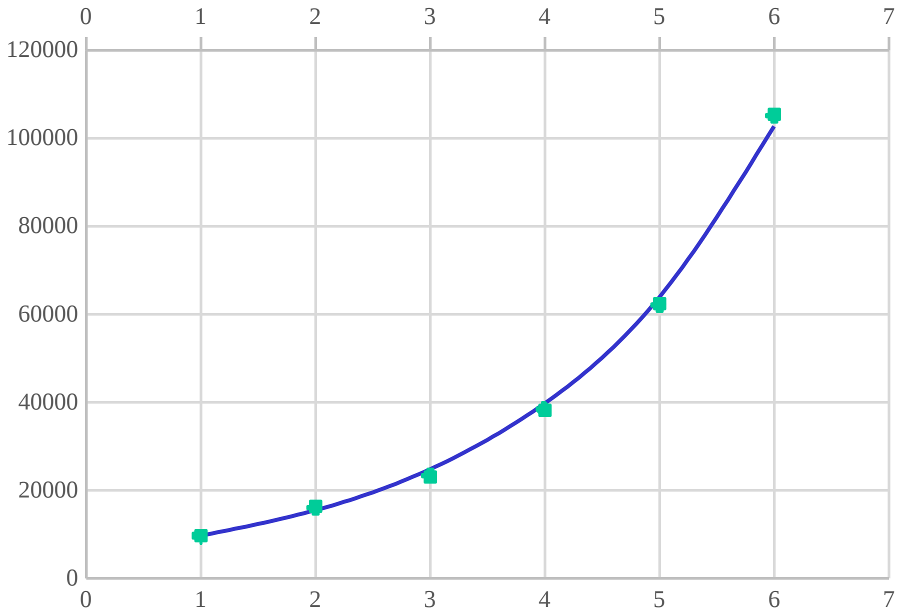
$a' = 8.68$ $a = e^{a'} = 5{,}881$

$y = (5\ 881)\ 1.61^{x}$

Year (x_i)	Number of transistors (y_i)
1	9500
2	16000
3	23000
4	38000
5	62000
6	105000

Example of a nonlinear model

Example 1: Estimated number of transistors in the following 6 years



Year (x_i)	Number of transistors (y_i)
1	9500
2	16000
3	23000
4	38000
5	62000
6	105000

$$y = (5\ 881)\ 1.61^x$$

Example of linear regression

In R (linear regression without transformation)

```
> D = read.table("regr1.in",header=TRUE)
> lr.out = lm(D$number~D$year)
> summary(lr.out)

Call:
lm(formula = D$number ~ D$year)

Residuals:
       1        2        3        4        5        6 
 12285.7    771.4 -10242.9 -13257.1  -7271.4  17714.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -20800      13156  -1.581  0.18904   
D$year         18014       3378   5.332  0.00596 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 14130 on 4 degrees of freedom
Multiple R-squared: 0.8767,   Adjusted R-squared: 0.8458 
F-statistic: 28.44 on 1 and 4 DF,  p-value: 0.005955
```

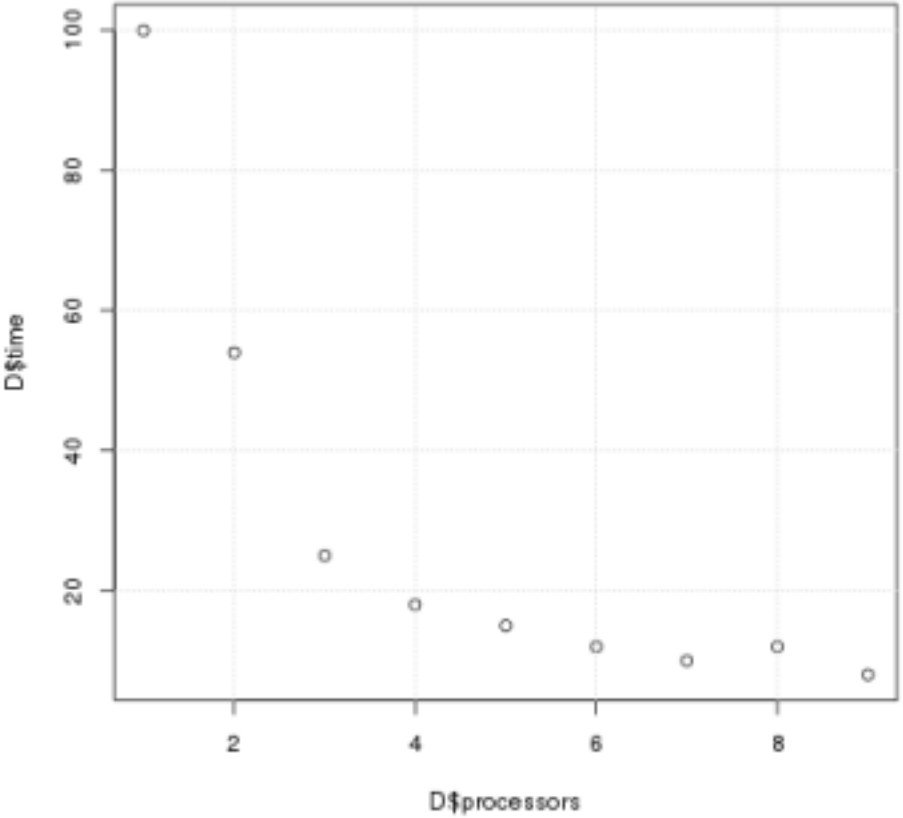
Example of linear regression

In R (linear regression with log transformation)

```
> D = read.table("regr1.in",header=TRUE)
> lr.out = lm(log(D$number)~D$year)
> summary(lr.out)

Call:
lm(formula = log(D$number) ~ D$year)

Residuals:
        1         2         3         4         5         6 
 0.005835  0.053444 -0.057338 -0.028934 -0.013073  0.040065 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.67952    0.04364  198.87 3.84e-09 ***
D$year       0.47369    0.01121   42.27 1.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04688 on 4 degrees of freedom
Multiple R-squared: 0.9978,   Adjusted R-squared: 0.9972 
F-statistic:  1787 on 1 and 4 DF,  p-value: 1.873e-06
```

Example of a nonlinear model

Example 2: CPU-time dependent of the number of processors



Processors (x_i)	CPU-Time (y_i)
1	100
2	54
3	25
4	18
5	15
6	12
7	10
8	12
9	8

$$y = a + b\, x$$

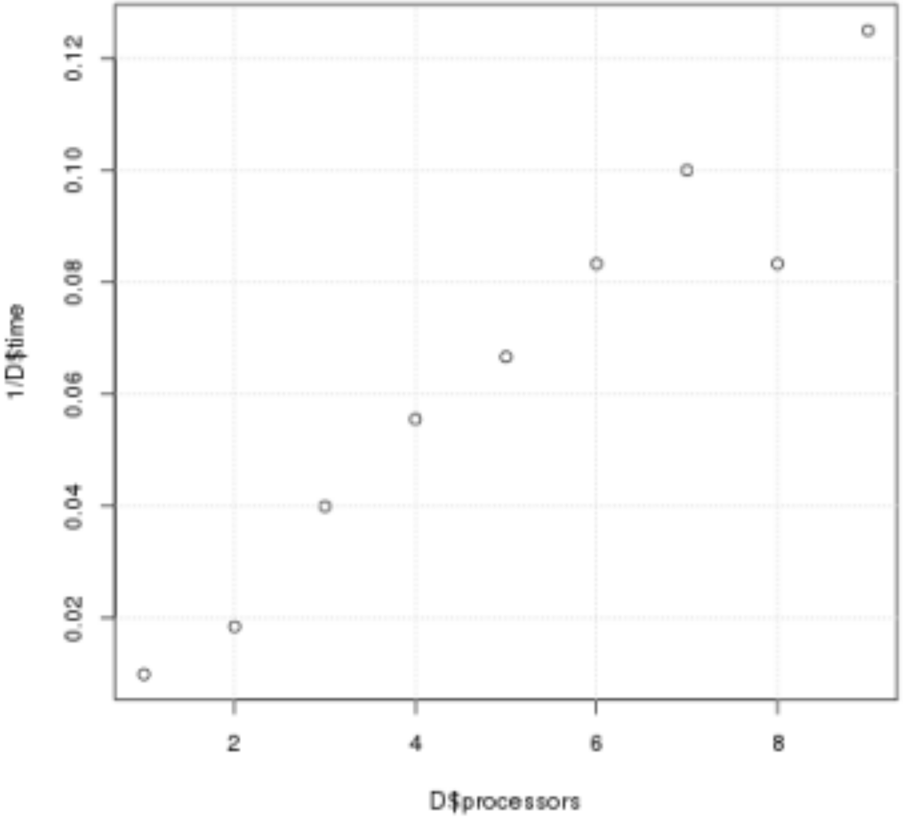
Example of linear regression

In R (linear regression without transformation)

```
> D = read.table("regr3.in",header=TRUE
> lr.out = lm(D$time~D$processors)
> summary(lr.out)

Call:
lm(formula = D$time ~ D$processors)

Residuals:
    Min      1Q  Median      3Q     Max
-20.889 -13.222  -0.722  10.278  36.444

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    72.389     14.246   5.081  0.00143 **
D$processors   -8.833      2.532  -3.489  0.01014 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.61 on 7 degrees of freedom
Multiple R-squared: 0.6349,   Adjusted R-squared: 0.5828
F-statistic: 12.17 on 1 and 7 DF,  p-value: 0.01014
```

Example of a nonlinear model

Reciprocal transformation



Size (x_i)	CPU-Time ($_i$)
1	0.01
2	0.02
3	0.04
4	0.06
5	0.07
6	0.08
7	0.10
8	0.08
9	0.13

$$1/y = a + b\,x$$

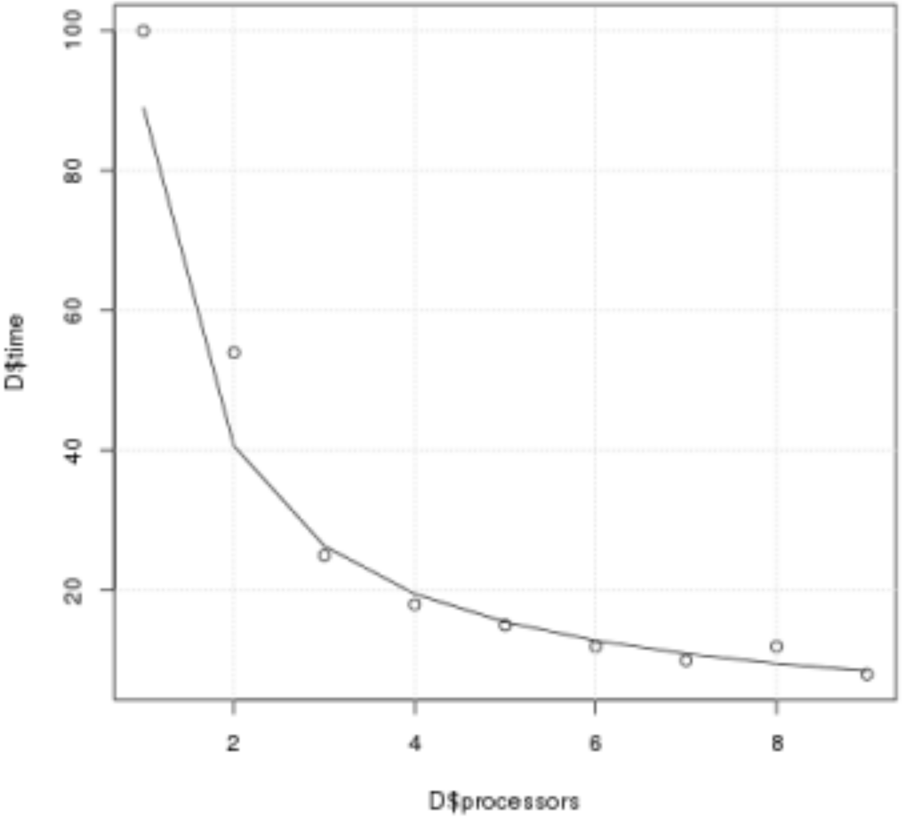
Example of linear regression

In R (linear regression with transformation)

```
> D = read.table("regr3.in",header=TRUE)
> lr.out = lm(1/D$time~D$processors)
> summary(lr.out)

Call:
lm(formula = 1/D$time ~ D$processors)

Residuals:
      Min         1Q    Median         3Q        Max
-0.021490 -0.001231  0.002029  0.005251  0.008547

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.002140   0.007123   -0.30    0.773
D$processors  0.013370   0.001266   10.56 1.49e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009805 on 7 degrees of freedom
Multiple R-squared: 0.941,    Adjusted R-squared: 0.9325
F-statistic: 111.6 on 1 and 7 DF,  p-value: 1.49e-05
```

Example of a nonlinear model

Reciprocal transformation

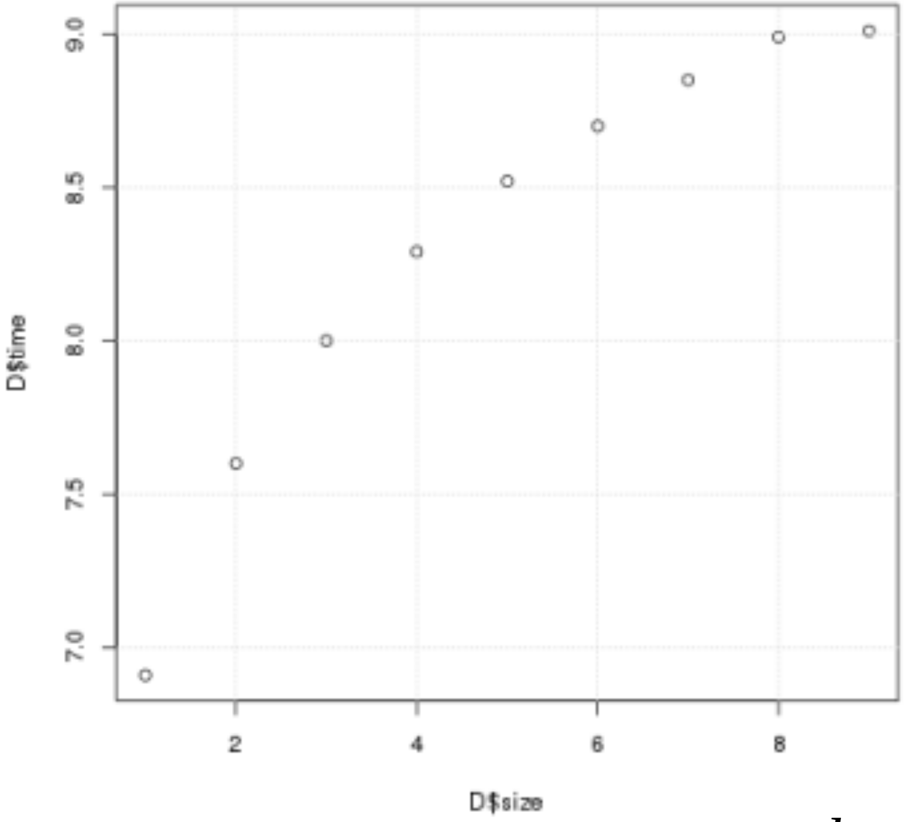


Processors (x_i)	CPU-Time (y_i)
1	100
2	54
3	25
4	18
5	15
6	12
7	10
8	12
9	8

$$y = 1 / (-0.002 + 0.013\, x)$$

Example of a nonlinear model

Example 3: CPU-time of binary search depending of the number of elements.



Size (x_i)	CPU-Time (y_i)
1	6.91
2	7.6
3	8.0
4	8.29
5	8.52
6	8.70
7	8.85
8	8.99
9	9.01

$$y = a + b\,x$$

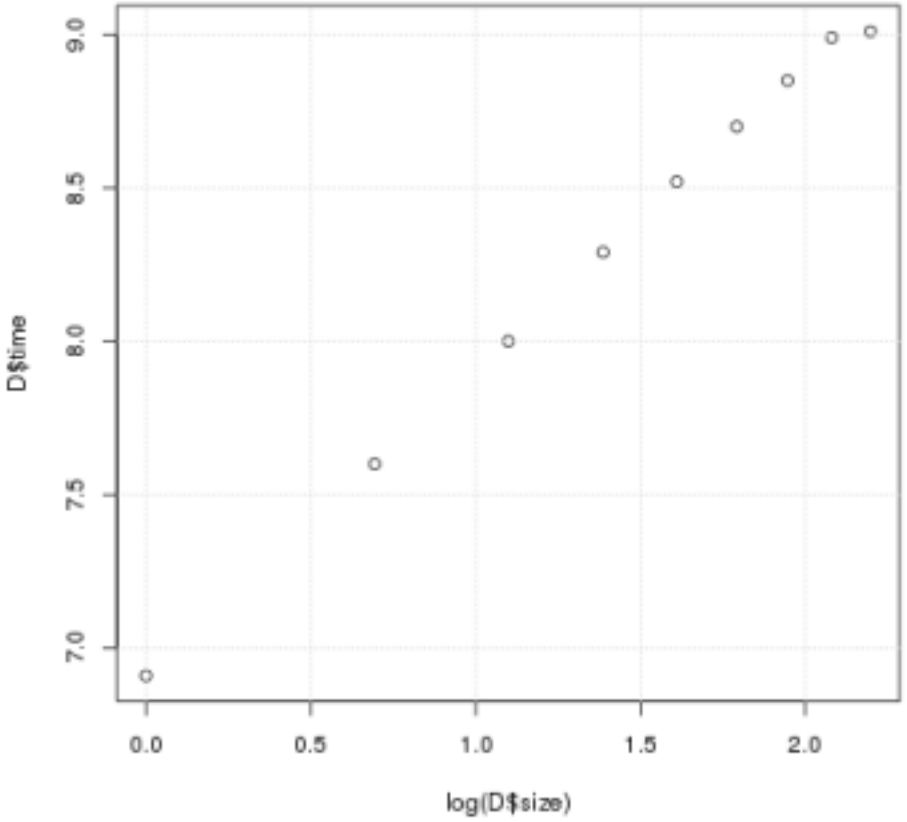
Example of linear regression

In R (linear regression without transformation)

```
> D = read.table("regr4.in",header=TRUE
> lr.out = lm(D$time~D$size)
> summary(lr.out)

Call:
lm(formula = D$time ~ D$size)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43022 -0.06289  0.04178  0.17044  0.21578 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.09556    0.17547  40.437 1.47e-09 ***
D$size       0.24467    0.03118   7.846 0.000103 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.2415 on 7 degrees of freedom
Multiple R-squared: 0.8979,   Adjusted R-squared: 0.8833 
F-statistic: 61.56 on 1 and 7 DF,  p-value: 0.0001031
```

Example of a nonlinear model

Logarithmic transformation



Size ($\log(x_i)$)	CPU-Time (y_i)
0	6.91
0.69	7.6
1.10	8.0
1.39	8.29
1.61	8.52
1.79	8.70
1.95	8.85
2.08	8.99
2.20	9.01

$$y = a + b \log(x)$$

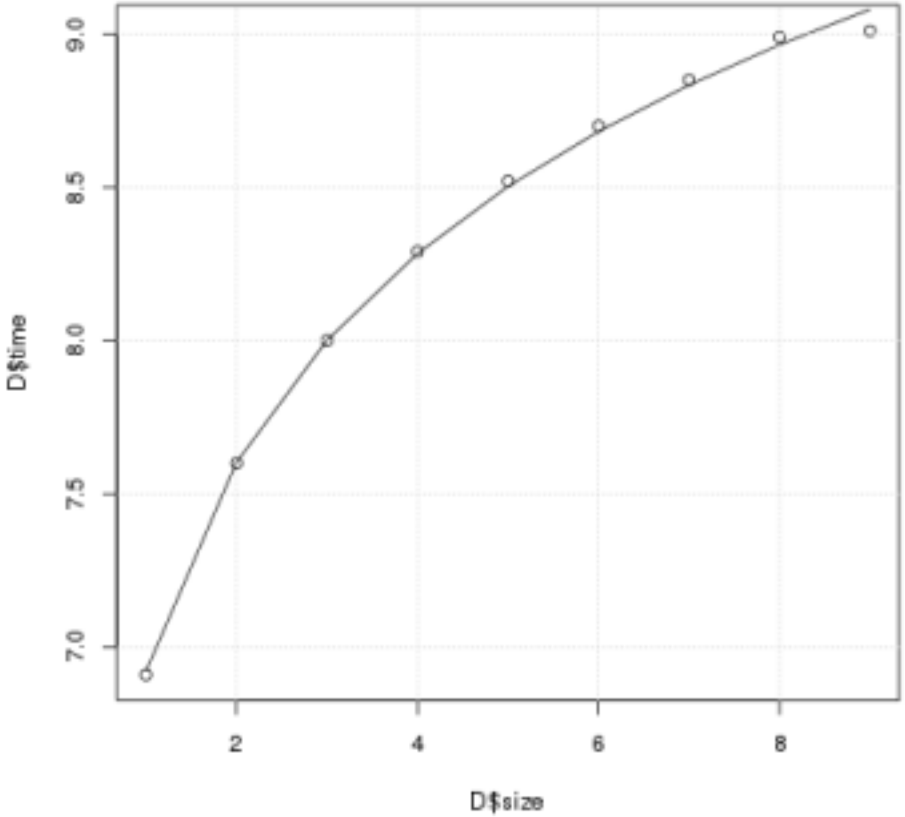
Example of linear regression

In R (linear regression with transformation)

```
> D = read.table("regr4.in",header=TRUE)
> lr.out = lm(D$time~log(D$size))
> summary(lr.out)

Call:
lm(formula = D$time ~ log(D$size))

Residuals:
      Min        1Q    Median        3Q       Max
-0.069968 -0.002525  0.006602  0.017410  0.025730

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.92165    0.02391  289.53 1.55e-15 ***
log(D$size)  0.98229    0.01517   64.75 5.51e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03086 on 7 degrees of freedom
Multiple R-squared: 0.9983,   Adjusted R-squared: 0.9981
F-statistic:  4192 on 1 and 7 DF,  p-value: 5.508e-11
```

Example of a nonlinear model

Logarithm transformation

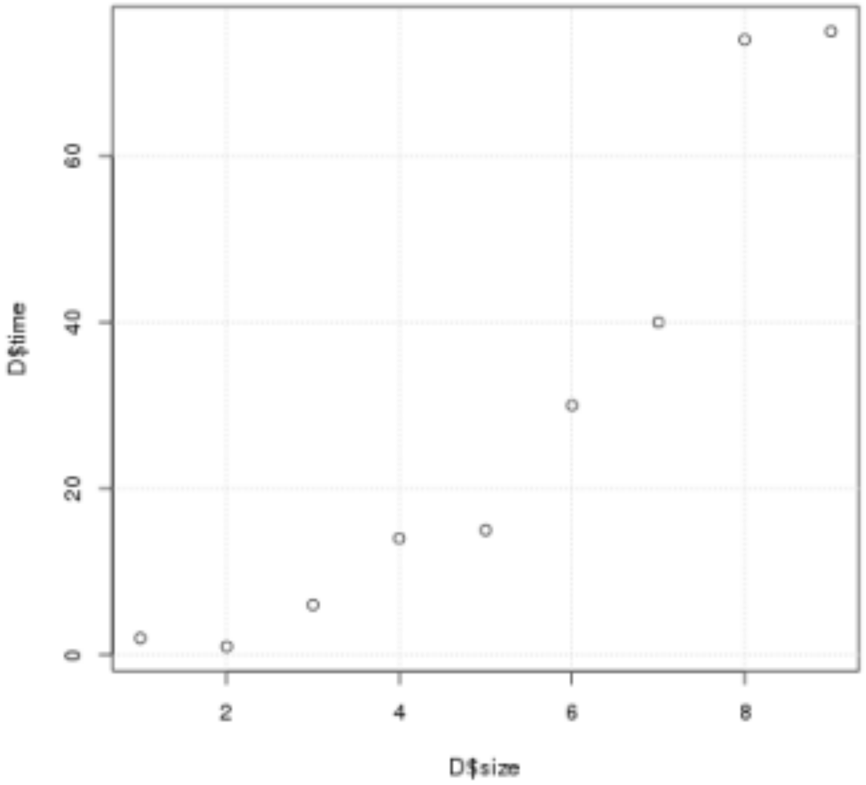


Size (x_i)	CPU-Time (y_i)
1	6.91
2	7.6
3	8.0
4	8.29
5	8.52
6	8.70
7	8.85
8	8.99
9	9.01

$$y = 6.92 + 0.98 \log(x)$$

Example of a nonlinear model

Example 4: CPU-time dependent of instance size



Size (x_i)	CPU-Time (y_i)
1	2
2	1
3	6
4	14
5	15
6	30
7	40
8	74
9	75

$$y = a + b\,x$$

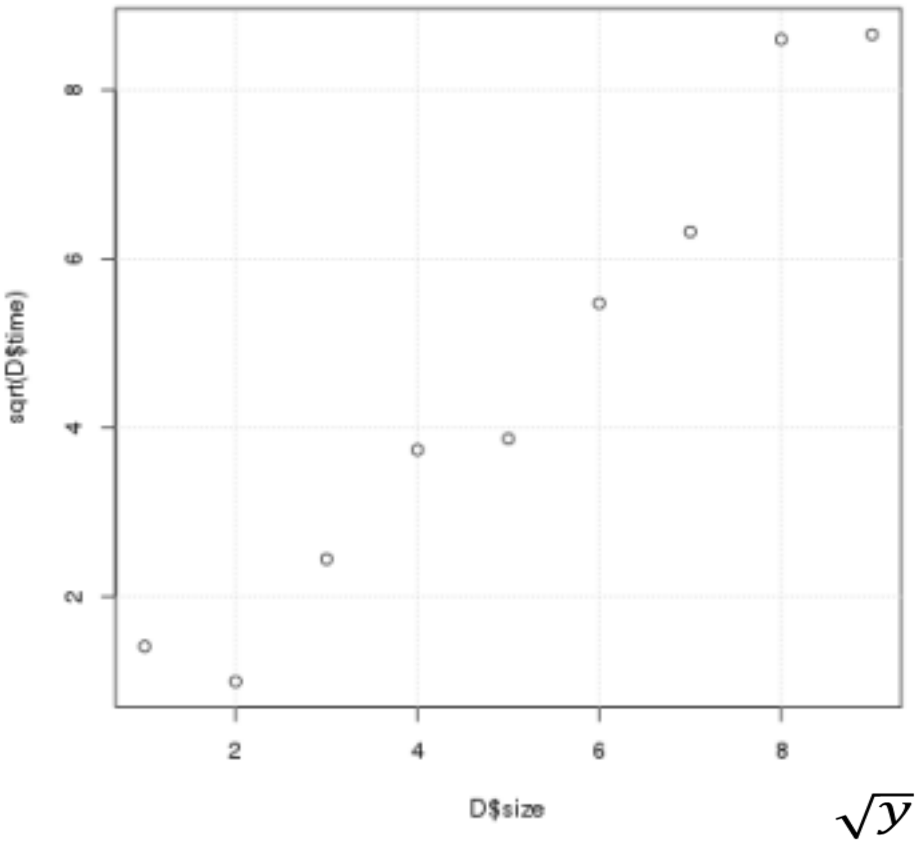
Example of linear regression

In R (linear regression without transformation)

```
> D = read.table("regr2.in",header=TRUE)
> lr.out = lm(D$time~D$size)
> summary(lr.out)

Call:
lm(formula = D$time ~ D$size)

Residuals:
    Min      1Q  Median      3Q     Max
-13.556  -8.389  -2.722   6.778  15.694

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21.028      7.881  -2.668 0.032087 *
D$size         9.917      1.401   7.081 0.000197 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.85 on 7 degrees of freedom
Multiple R-squared: 0.8775,   Adjusted R-squared:  0.86
F-statistic: 50.14 on 1 and 7 DF,  p-value: 0.000197
```

Example of a nonlinear model

Square root transformation of the dependent variable



Size (x_i)	CPU-Time ($_i$)
1	1.41
2	1
3	2.45
4	3.74
5	3.87
6	5.48
7	6.32
8	8.60
9	8.66

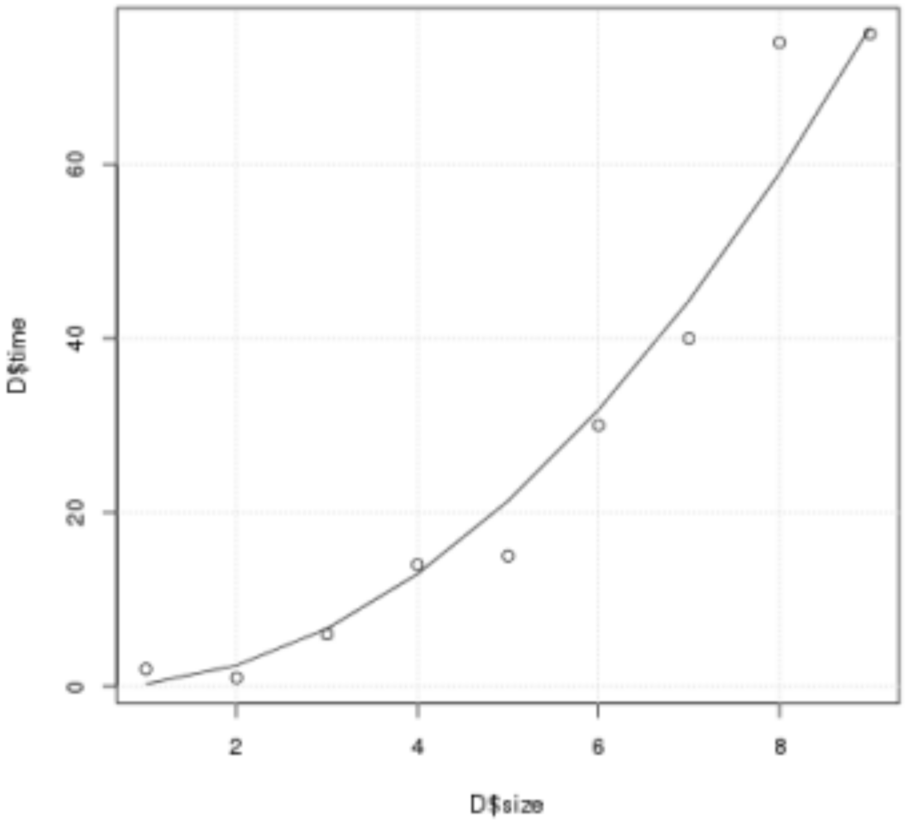
$$\sqrt{y} = a + b\, x$$

Example of linear regression

In R (linear regression with transformation)

```
> D = read.table("regr2.in",header=TRUE)
> lr.out = lm(sqrt(D$time)~D$size)
> summary(lr.out)

Call:
lm(formula = sqrt(D$time) ~ D$size)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7429 -0.3339 -0.1238  0.1471  0.9226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.49055    0.44817  -1.095     0.31
D$size       1.02128    0.07964  12.823 4.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6169 on 7 degrees of freedom
Multiple R-squared: 0.9592,   Adjusted R-squared: 0.9533
F-statistic: 164.4 on 1 and 7 DF,  p-value: 4.068e-06
```

Example of a nonlinear model



Size (x_i)	CPU-Time (y_i)
1	2
2	1
3	6
4	14
5	15
6	30
7	40
8	74
9	75

$$y = (-0.49 + 1.02\,x)^2$$

Summary of transformations

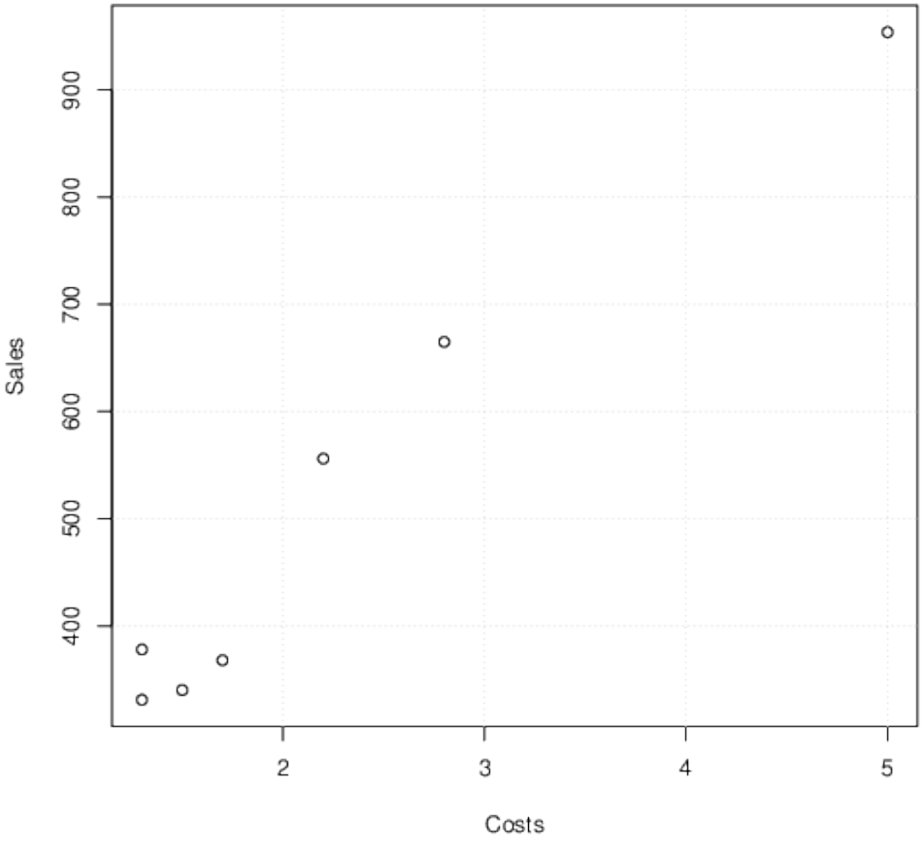
• Standard linear regression:	$y = a + bx$	$\hat{y} = a + bx$
• Exponential model :	$\ln(y) = a + bx$	$\hat{y} = e^{a + bx}$
• Quadratic model :	$\sqrt{y} = a + bx$	$\hat{y} = (a + bx)^2$
• Reciprocal model	$1/y = a + bx$	$\hat{y} = 1 / (a + bx)$
• Logarithmic model :	$y = a + b \ln(x)$	$\hat{y} = a + b \ln(x)$
• Power model :	$\ln(y) = \ln(a) + b \ln(x)$	$\hat{y} = ax^b$

Assumptions

- Linear relation between the independent and the dependent variable
- Independence of residuals
- Normal distribution of residuals
- Equal variance of residuals

Assumptions

Example: Monthly E-Commerce Sales and On-line advertising costs



Costs	Sales
1.7	368
1.5	340
2.8	665
5.0	954
1.3	331
2.2	556
1.3	376

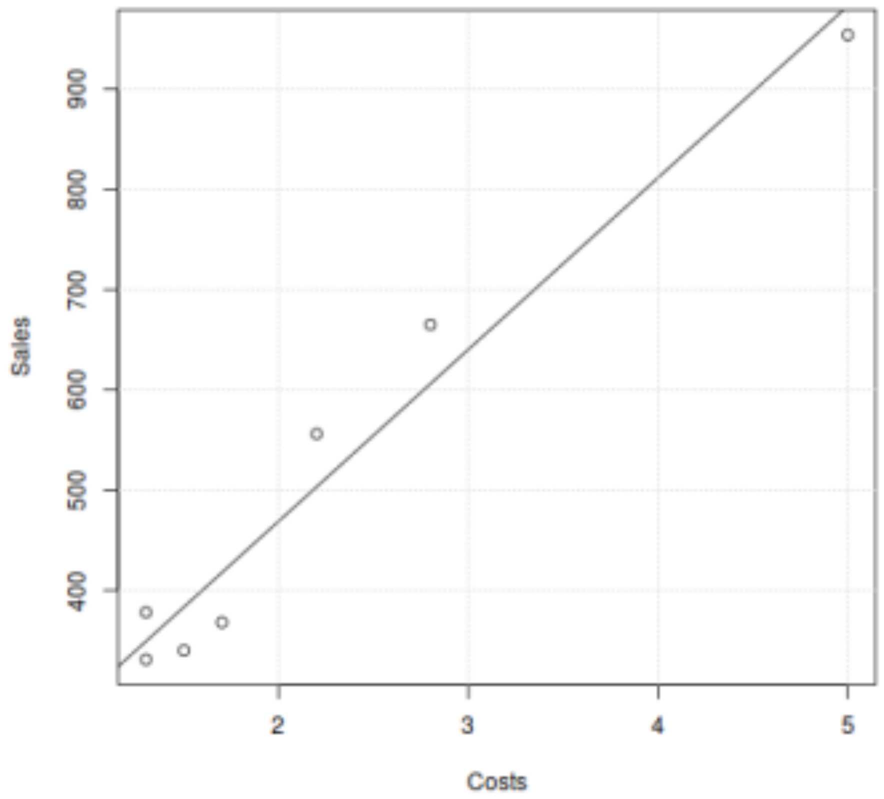
Assumptions

In R (linear regression)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   126.53      39.72   3.185 0.024393 *
R$Cost        171.28      15.46  11.077 0.000104 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.19 on 5 degrees of freedom
Multiple R-squared:  0.9608,  Adjusted R-squared:  0.953
F-statistic: 122.7 on 1 and 5 DF,  p-value: 0.0001045
```

Assumptions

Example: Monthly E-Commerce Sales and On-line advertising costs

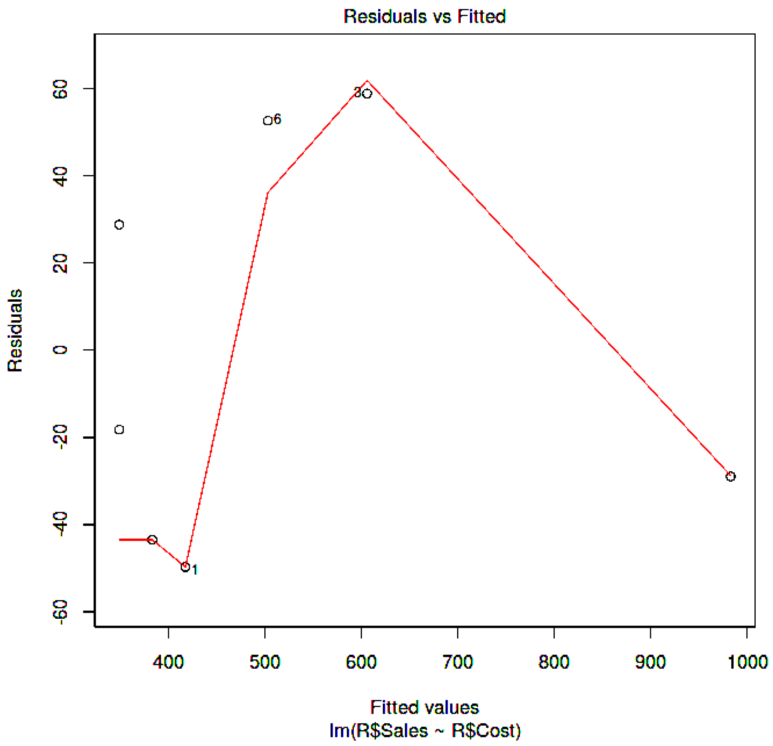


Costs	Sales
1.7	368
1.5	340
2.8	665
5.0	954
1.3	331
2.2	556
1.3	376

Assumptions

Linear relation between the independent and the dependent variable

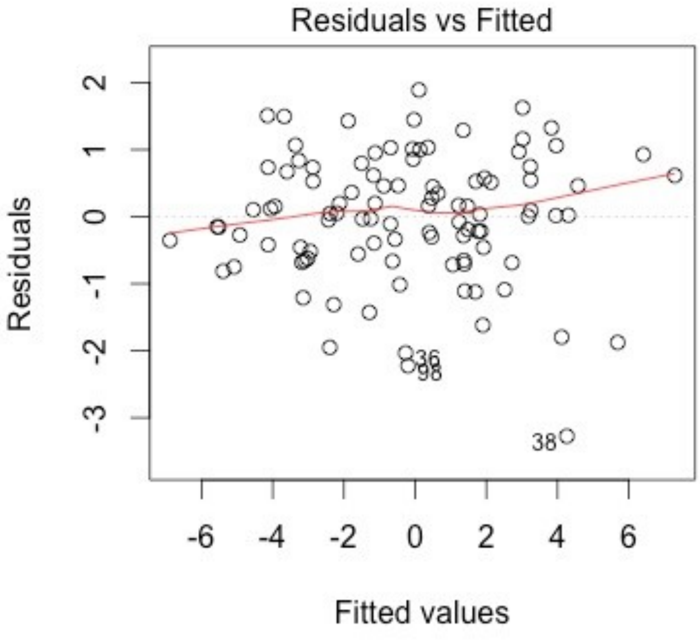
```
> plot(lm(R$Sales~R$Cost))
```



It suggests non-linearity

Plot of residuals versus fitted values

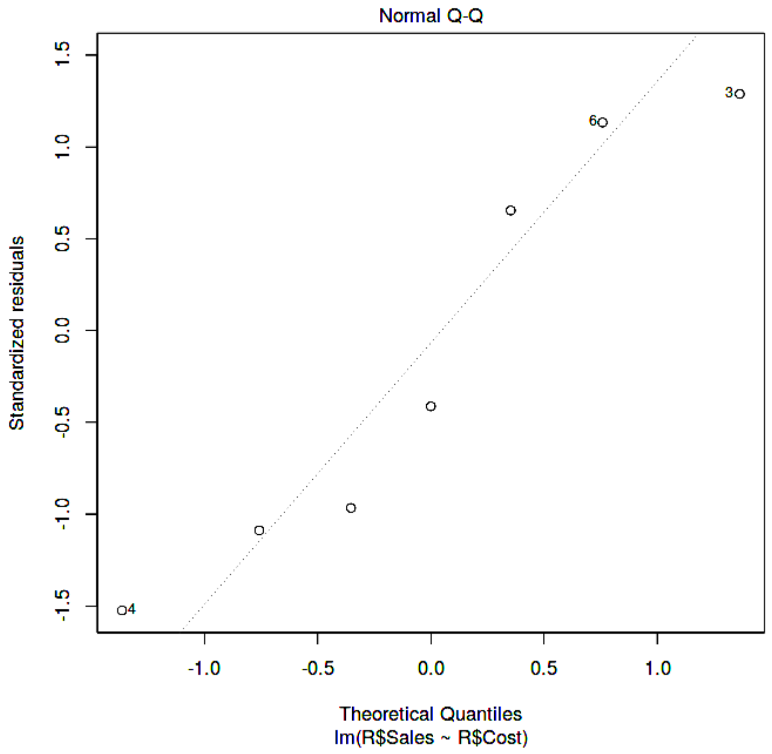
- The red line should be almost flat



Assumptions

Normal distribution of residuals

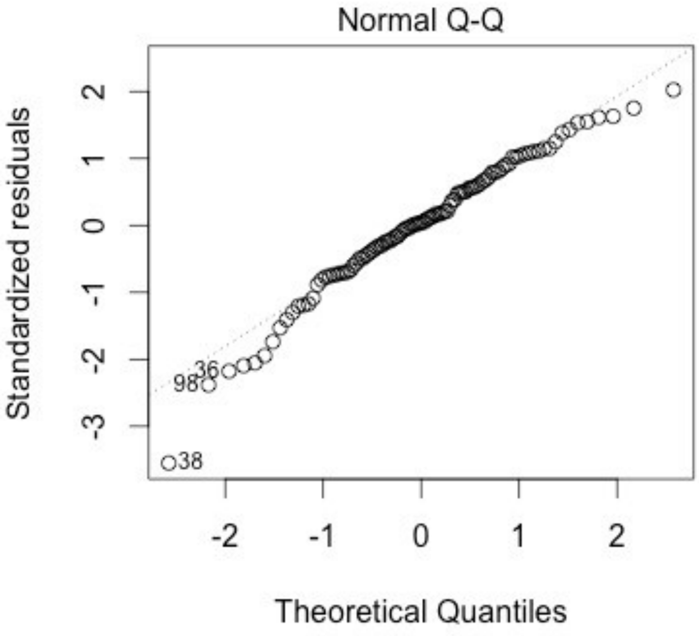
```
> plot(lm(R$Sales~R$Cost))
```



It suggests (close to) normality

Normal Q-Q plot

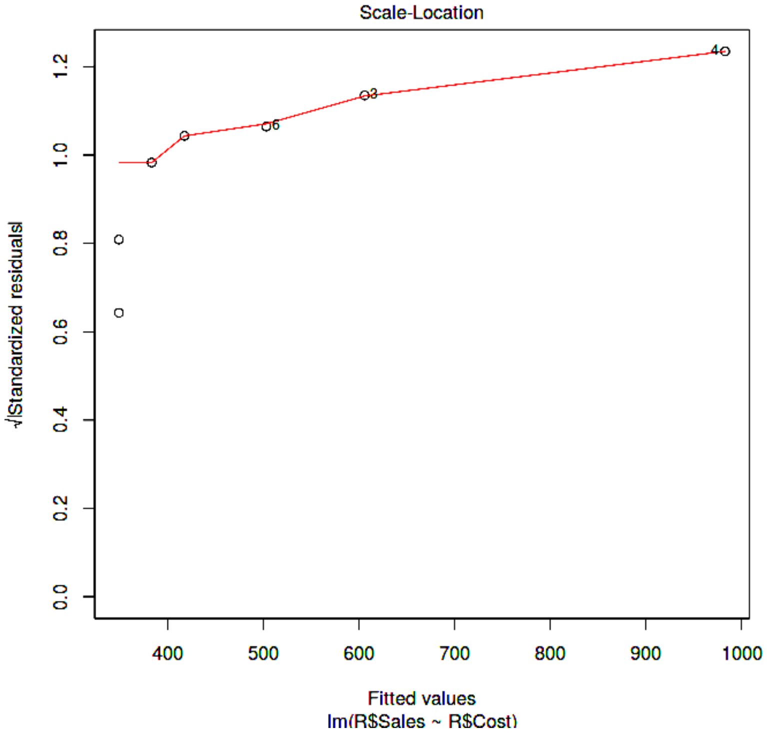
- The points should follow the line



Assumptions

Equal variance of residuals

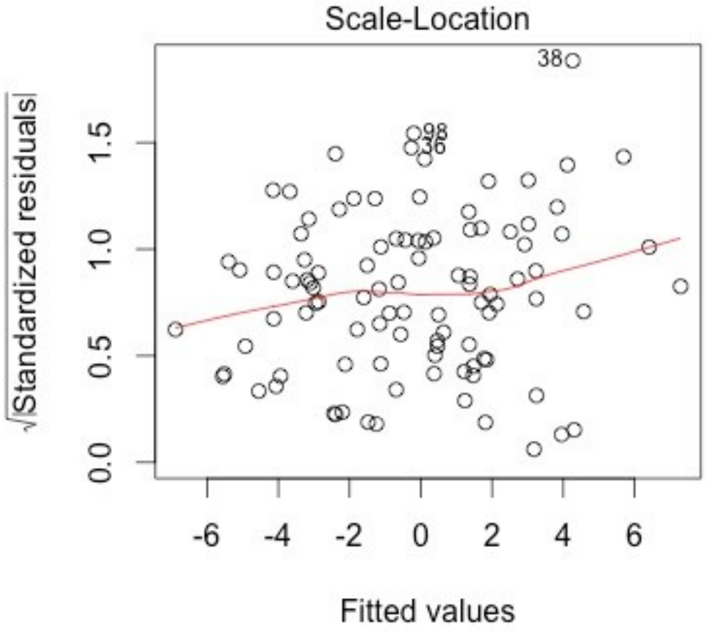
```
> plot(lm(R$Sales~R$Cost))
```



It suggests unequal variance of residuals

Scale-location plot

- Horizontal line with equally spread points



Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
38
98
36
-6
-4
-2
0
2
4
6
Fitted values