Workshop 7.2a: Introduction to Linear models

Murray Logan

July 19, 2017

Table of contents

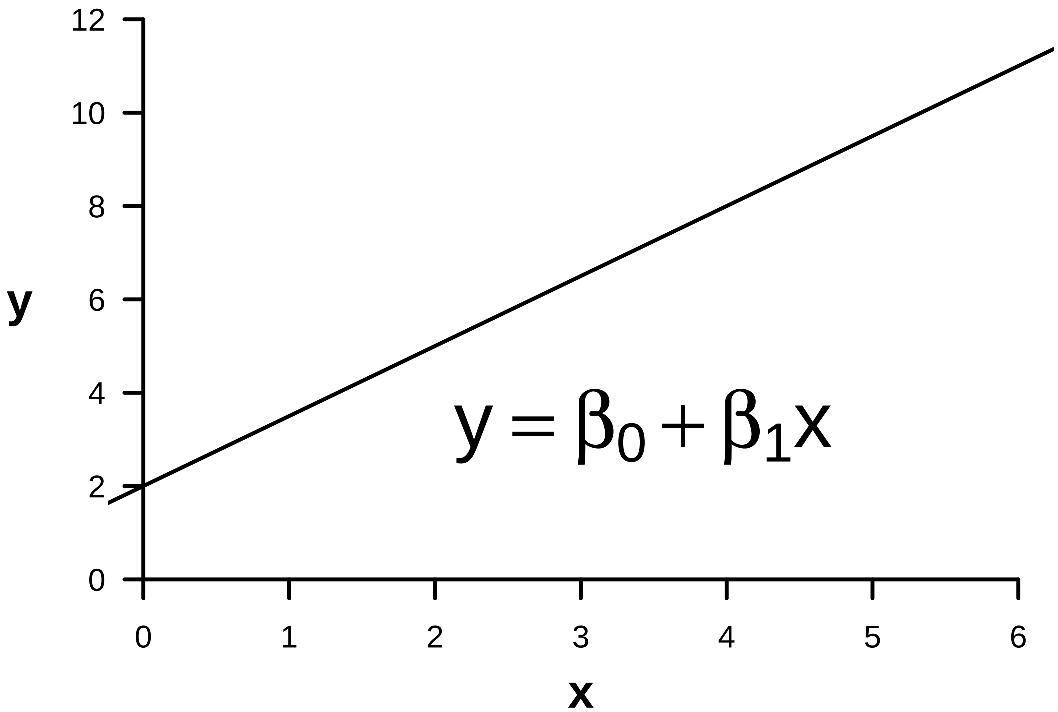
1. Revision

1.1. Aims of statistical modelling

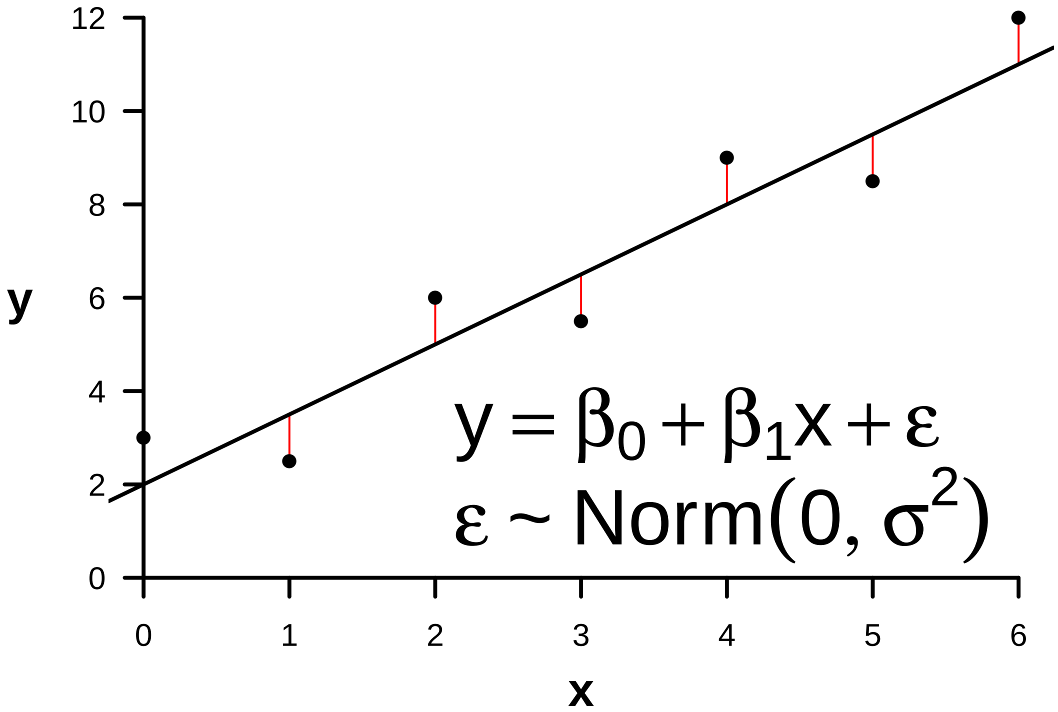
Use samples to:

- Describe relationships
- Inference testing (relationships/effects)
- Predictive models

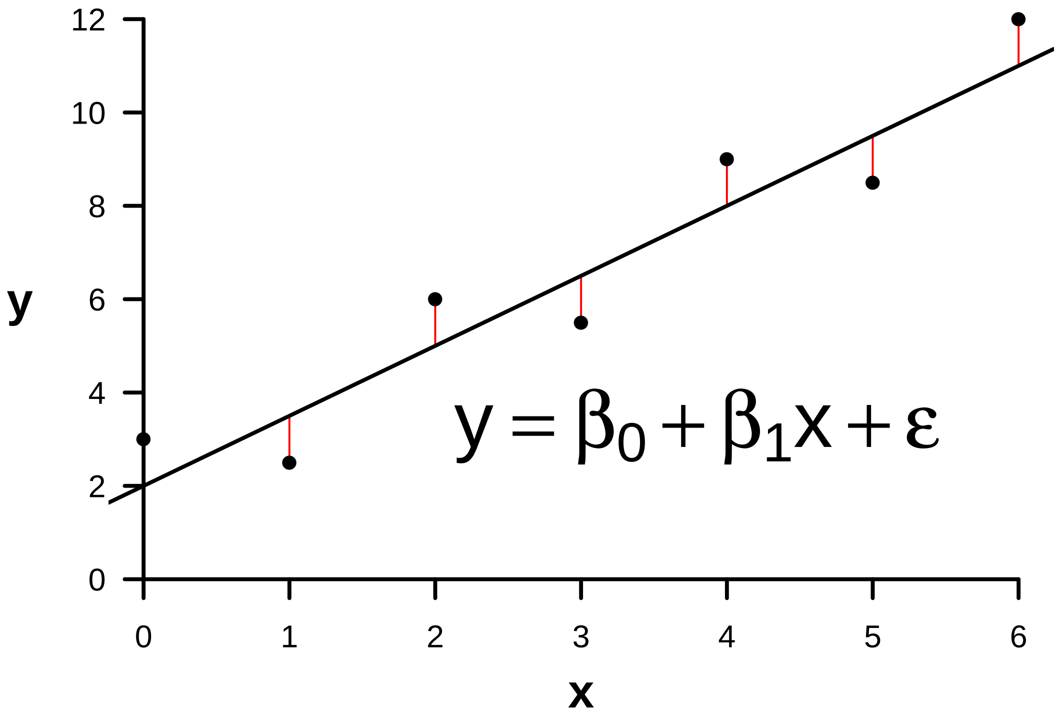
1.2. Mathematical models



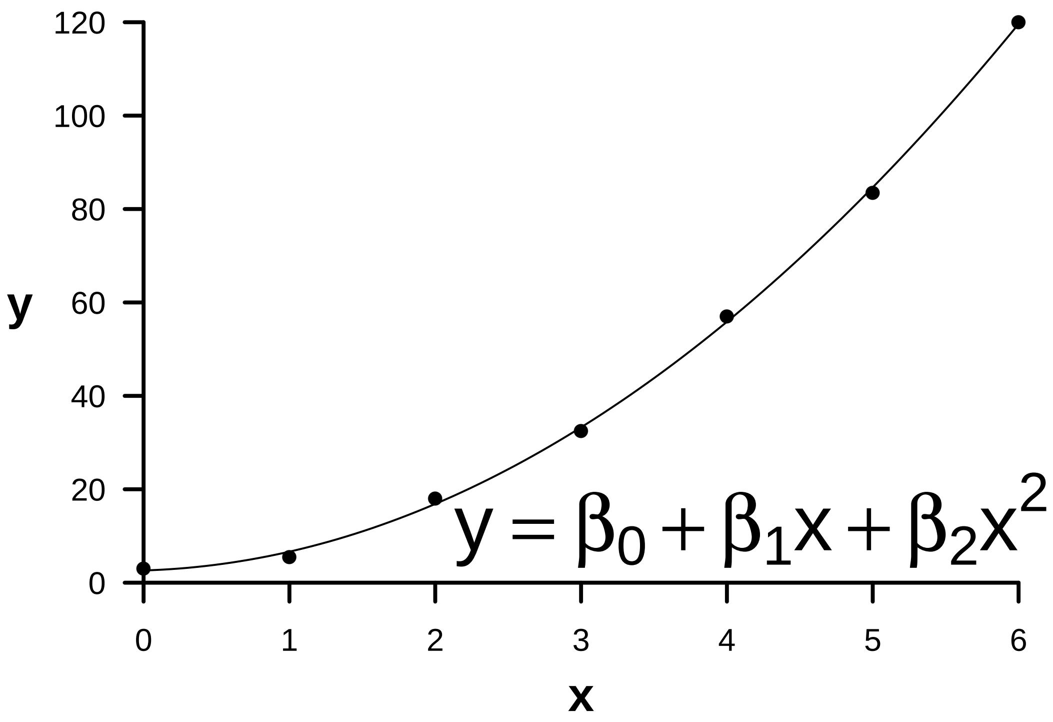
1.3. Statistical models



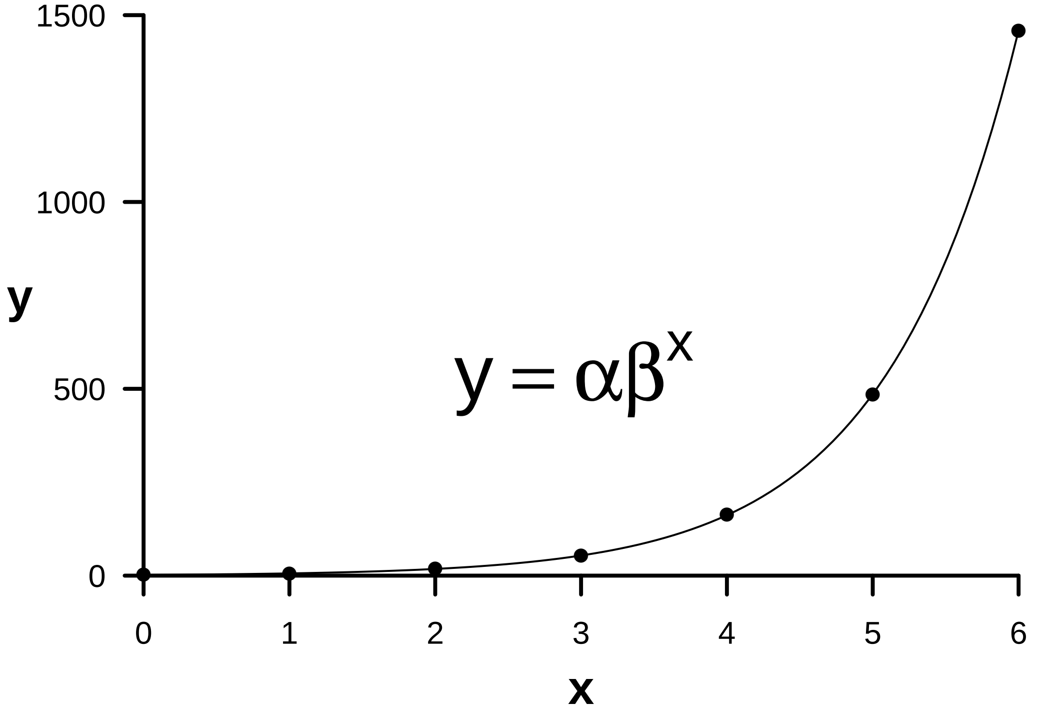
1.4. Linear models



1.5. Linear models



1.6. Non-linear models



1.7. Linear models

$$y_i = \quad \beta_0 \quad + \quad \beta_1 \quad \times \quad x_1 \quad + \quad \epsilon_1$$

$$\underset{\textit{variable}}{\textit{response}} = \underbrace{\underbrace{\underset{\textit{intercept}}{\textit{population}}}_{\text{intercept term}} + \underbrace{\underset{\textit{slope}}{\textit{population}} \times \underset{\textit{variable}}{\textit{predictor}}}_{\text{slope term}}}_{\text{Systematic component}} + \underbrace{\textit{error}}_{\text{Stoichastic component}}$$

1.8. Linear models

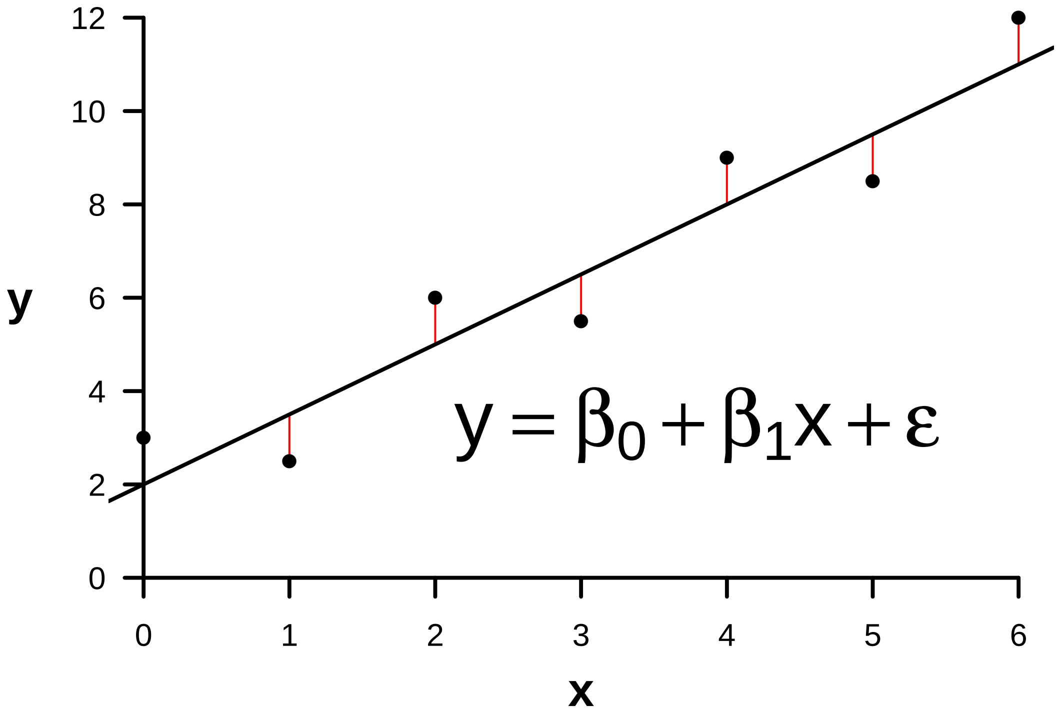
$$y_i = \quad \beta_0 \quad + \quad \beta_1 \quad \times \quad x_1 \quad + \quad \varepsilon_1$$

$$\underset{\textbf{vector}}{\textit{response}} = \underbrace{\underbrace{\underset{\textbf{single value}}{\textit{intercept}}}_{\text{intercept term}} + \underbrace{\underset{\textbf{single value}}{\textit{slope}} \times \underset{\textbf{vector}}{\textit{predictor}}}_{\text{slope term}}}_{\text{Systematic component}} + \underbrace{\textit{error}}_{\text{Stoichastic component}}$$

1.9. Vectors and Matrices

Vector	Matrix
$\begin{pmatrix} 3.0 \\ 2.5 \\ 6.0 \\ 5.5 \\ 9.0 \\ 8.6 \\ 12.0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{pmatrix}$
Has length ONLY	Has length AND width

1.10. Estimation



Ordinary Least Squares

1.11. Estimation

Y	X
3	0
2.5	1
6	2
5.5	3
9	4
8.6	5
12	6

$$
\begin{array}{ccccccc}
3.0 & = & \beta_0 \times 1 & + & \beta_1 \times 0 & + & \varepsilon_1 \\
2.5 & = & \beta_0 \times 1 & + & \beta_1 \times 1 & + & \varepsilon_1 \\
6.0 & = & \beta_0 \times 1 & + & \beta_1 \times 2 & + & \varepsilon_2 \\
5.5 & = & \beta_0 \times 1 & + & \beta_1 \times 3 & + & \varepsilon_3 \\
9.0 & = & \beta_0 \times 1 & + & \beta_1 \times 4 & + & \varepsilon_4 \\
8.6 & = & \beta_0 \times 1 & + & \beta_1 \times 5 & + & \varepsilon_5 \\
12.0 & = & \beta_0 \times 1 & + & \beta_1 \times 6 & + & \varepsilon_6
\end{array}
$$

1.12. Estimation

$$
\begin{array}{ccccccccc}
3.0 & = & \beta_0 \times 1 & + & \beta_1 \times 0 & + & \varepsilon_1 \\
2.5 & = & \beta_0 \times 1 & + & \beta_1 \times 1 & + & \varepsilon_1 \\
6.0 & = & \beta_0 \times 1 & + & \beta_1 \times 2 & + & \varepsilon_2 \\
5.5 & = & \beta_0 \times 1 & + & \beta_1 \times 3 & + & \varepsilon_3 \\
9.0 & = & \beta_0 \times 1 & + & \beta_1 \times 4 & + & \varepsilon_4 \\
8.6 & = & \beta_0 \times 1 & + & \beta_1 \times 5 & + & \varepsilon_5 \\
12.0 & = & \beta_0 \times 1 & + & \beta_1 \times 6 & + & \varepsilon_6
\end{array}
$$

$$
\underbrace{\begin{pmatrix} 3.0 \\ 2.5 \\ 6.0 \\ 5.5 \\ 9.0 \\ 8.6 \\ 12.0 \end{pmatrix}}_{\text{Response values}} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{pmatrix}}_{\text{Model matrix}} \times \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\text{Parameter vector}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}}_{\text{Residual vector}}
$$

1.13. Inference testing

Ho:$\beta_1 = 0$ (slope equals zero)

The t-statistic

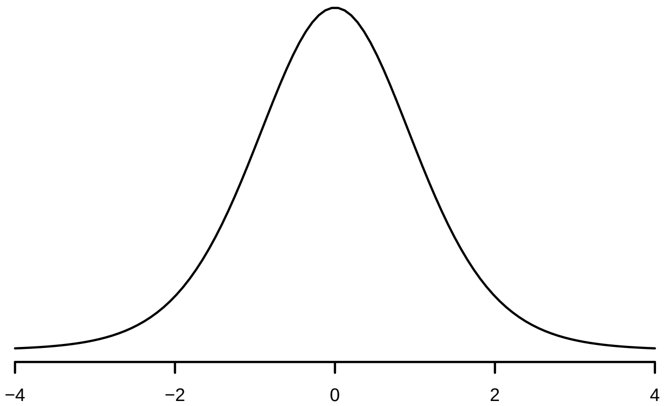
$$t = \frac{param}{SE_{param}}$$

$$t = \frac{\beta_1}{SE_{\beta_1}}$$

1.14. Inference testing

Ho:$\beta_1 = 0$ (slope equals zero)

The t-statistic and the t distribution



2. Linear model Assumptions

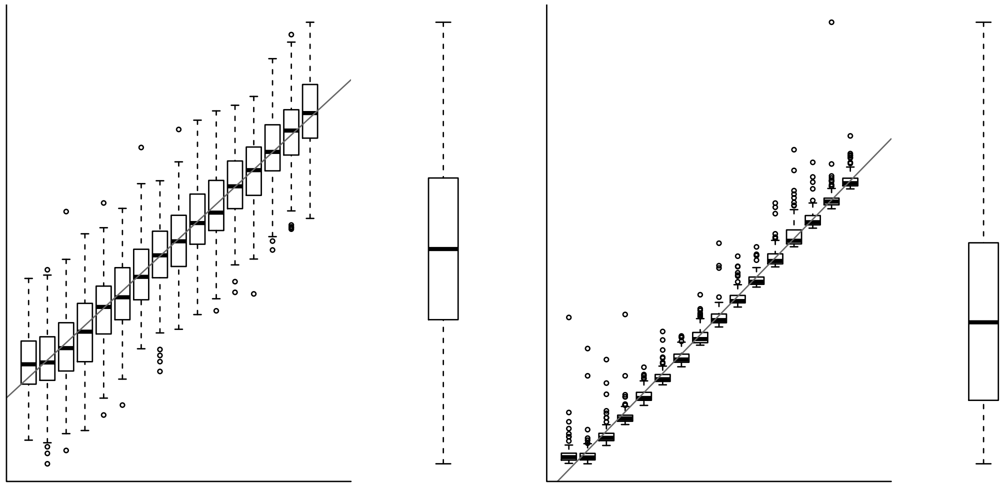
2.1. Assumptions

- Independence - unbiased, scale of treatment

- Normality - residuals
- Homogeneity of variance - residuals
- Linearity

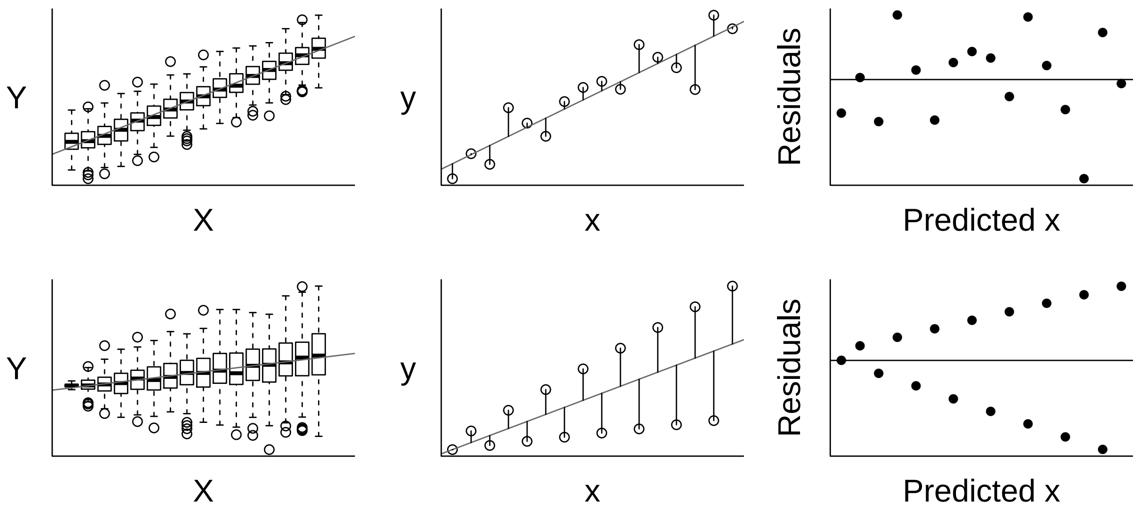
2.2. Assumptions

2.2.1. Normality



2.3. Assumptions

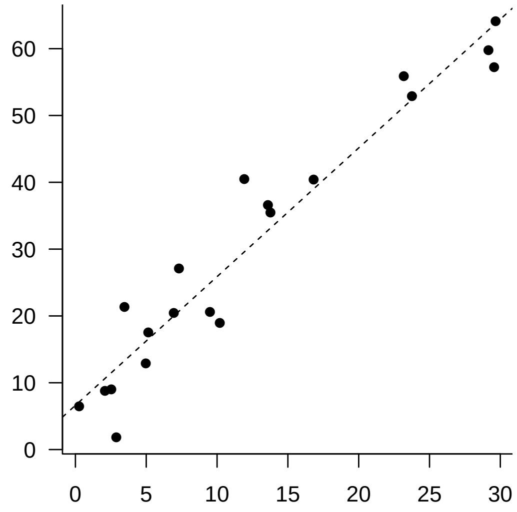
2.3.1. Homogeneity of variance

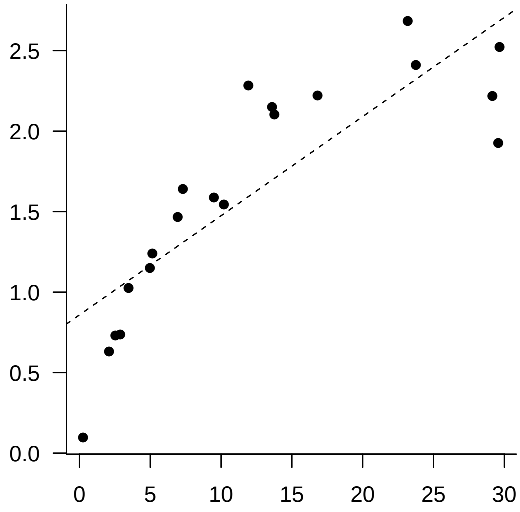


2.4. Assumptions

2.4.1. Linearity

Trendline

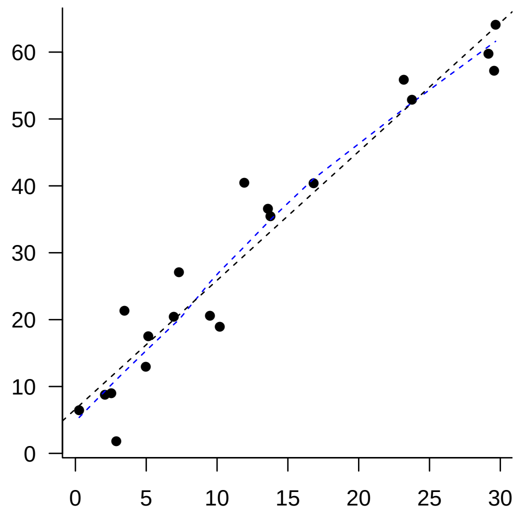


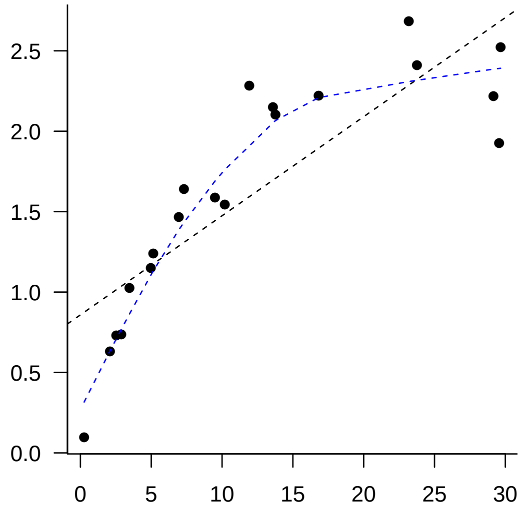


2.5. Assumptions

2.5.1. Linearity

Loess (lowess) smoother

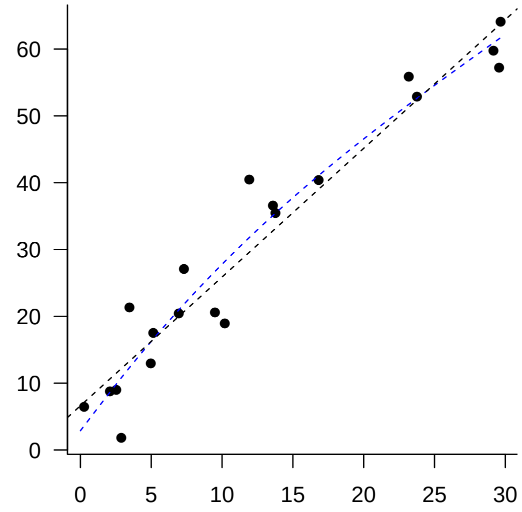


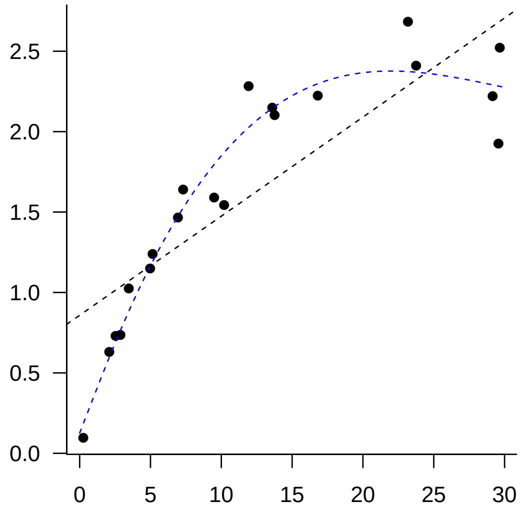


2.6. Assumptions

2.6.1. Linearity

Spline smoother





2.7. Assumptions

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

2.8. Assumptions

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

2.9. Example

Make these data and call the data frame `DATA`

Y	X
3	0
2.5	1
6	2
5.5	3
9	4
8.6	5
12	6

2.10. Example

Make these data and call the data frame `DATA`

Y	X
3	0
2.5	1
6	2
5.5	3
9	4
8.6	5
12	6

- try this...

```
> DATA <- data.frame(Y=c(3, 2.5, 6.0, 5.5, 9.0, 8.6, 12), X=0:6)
```

2.11. Worked Examples

```
> fert <- read.csv('../data/fertilizer.csv', strip.white=T)
> fert
```

```
   FERTILIZER YIELD
1          25    84
2          50    80
3          75    90
4         100   154
5         125   148
6         150   169
7         175   206
8         200   244
9         225   212
10        250   248
```

```
> head(fert)
```

```
  FERTILIZER YIELD
```

```
1          25    84
2          50    80
3          75    90
4         100   154
5         125   148
6         150   169
```

```
> summary(fert)
```

```
   FERTILIZER         YIELD
 Min.   : 25.00   Min.   : 80.0
 1st Qu.: 81.25   1st Qu.:104.5
 Median :137.50   Median :161.5
 Mean   :137.50   Mean   :163.5
 3rd Qu.:193.75   3rd Qu.:210.5
 Max.   :250.00   Max.   :248.0
```

```
> str(fert)
```

```
'data.frame':   10 obs. of  2 variables:
 $ FERTILIZER: int  25 50 75 100 125 150 175 200 225 250
 $ YIELD     : int  84 80 90 154 148 169 206 244 212 248
```

```
> library(INLA)
>
> fert.inla <- inla(YIELD ~ FERTILIZER, data=fert)
> summary(fert.inla)
```

Call: "inla(formula = YIELD ~ FERTILIZER, data = fert)"

Time used: Pre-processing Running inla Post-processing Total 0.3043 0.0715 0.0217 0.3974

Fixed effects: mean sd 0.025quant 0.5quant 0.975quant mode kld (Intercept) 51.9341 12.9747 25.9582 51.9335 77.8990 51.9339 0 FERTILIZER 0.8114 0.0836 0.6439 0.8114 0.9788 0.8114 0

The model has no random effects

Model hyperparameters: mean sd 0.025quant 0.5quant 0.975quant mode Precision for the Gaussian observations 0.0035 0.0015 0.0012 0.0032 0.007 0.0028

Expected number of effective parameters(std dev): 2.00(0.00) Number of equivalent replicates : 5.00

Marginal log-Likelihood: -61.65

2.12. Worked Examples

Question: is there a relationship between fertilizer concentration and grass yeild?

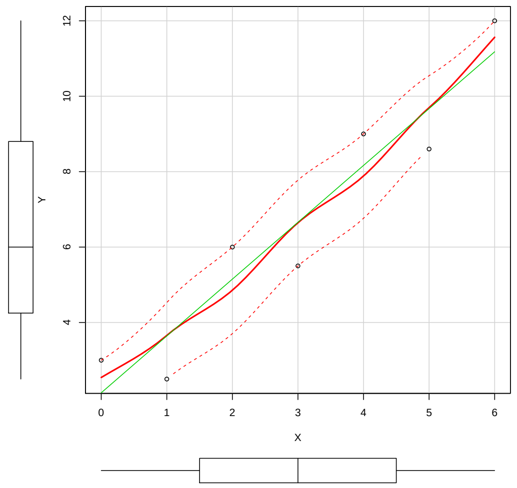
Linear model:

$$Y_i = \beta_0 + \beta_1 F_i + \varepsilon_i \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

2.13. Example

2.13.1. Exploratory data analysis

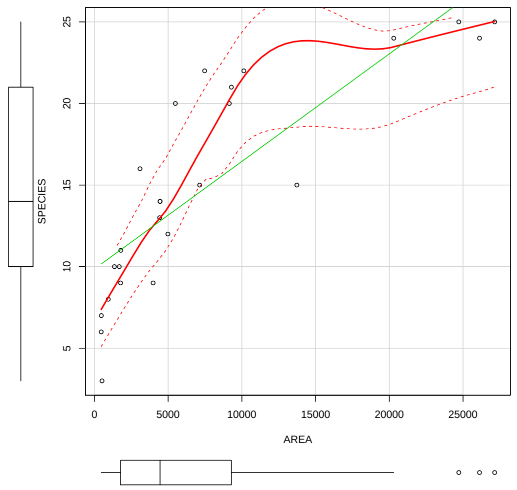
```
> library(car)
> scatterplot(Y~X, data=DATA)
```



2.14. Example

2.14.1. Exploratory data analysis

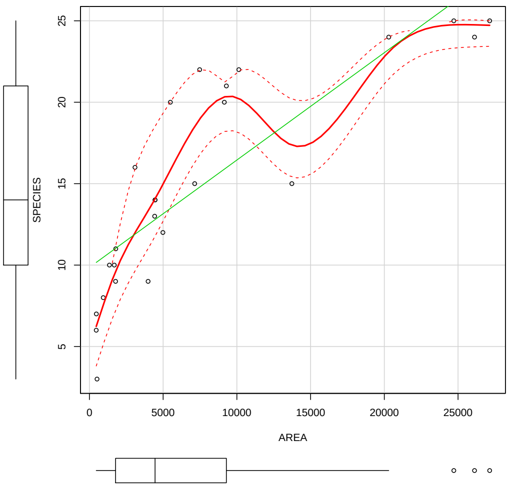
```
> library(car)
> peake <- read.csv('../data/peake.csv')
> scatterplot(SPECIES ~ AREA, data=peake)
```



2.15. Example

2.15.1. Exploratory data analysis

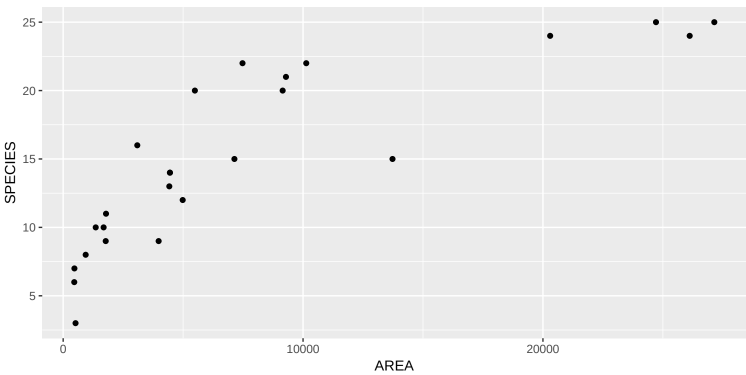
```
> scatterplot(SPECIES ~ AREA, data=peake,
+             smoother=gamLine)
```



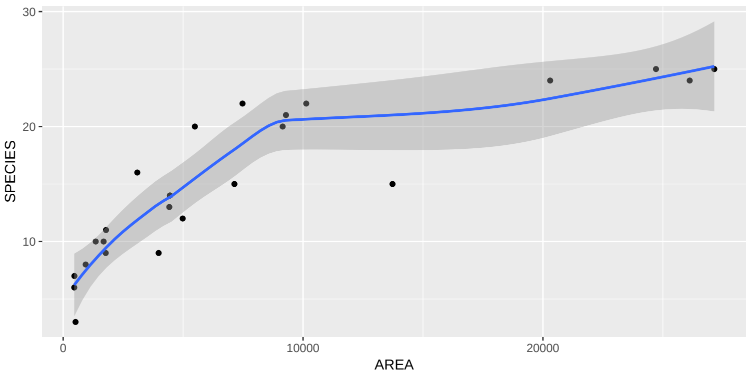
2.16. Example

2.16.1. Exploratory data analysis

```
> library(ggplot2)
> library(gridExtra)
> ggplot(peake, aes(y=SPECIES, x=AREA)) + geom_point()
```




```
> ggplot(peake, aes(y=SPECIES, x=AREA)) + geom_point() +
+     geom_smooth()
```




```
> p2 <- ggplot(peake, aes(y=SPECIES, x=1)) + geom_boxplot()
> p3 <- ggplot(peake, aes(y=AREA, x=1)) + geom_boxplot()
> grid.arrange(p1,p2,p3, ncol=3)
```

```
Error in arrangeGrob(...): object 'p1' not found
```

3. Simple Linear models in R

3.1. Linear models in R

```
> lm(formula, data= DATAFRAME)
```

Model	R formula	Description
$y_i = \beta_0 + \beta_1 x_i$	`y~1+x`	
	`y~x`	Full model
$y_i = \beta_0$	`y~1`	Null model
$y_i = \beta_1$	`y~-1+x`	Through origin

3.2. Example

3.2.1. Fit linear model

$$y_i = \beta_0 + \beta_1 x_i \qquad N(0, \sigma)$$

```
> DATA.lm<-lm(Y~X, data=DATA)
```

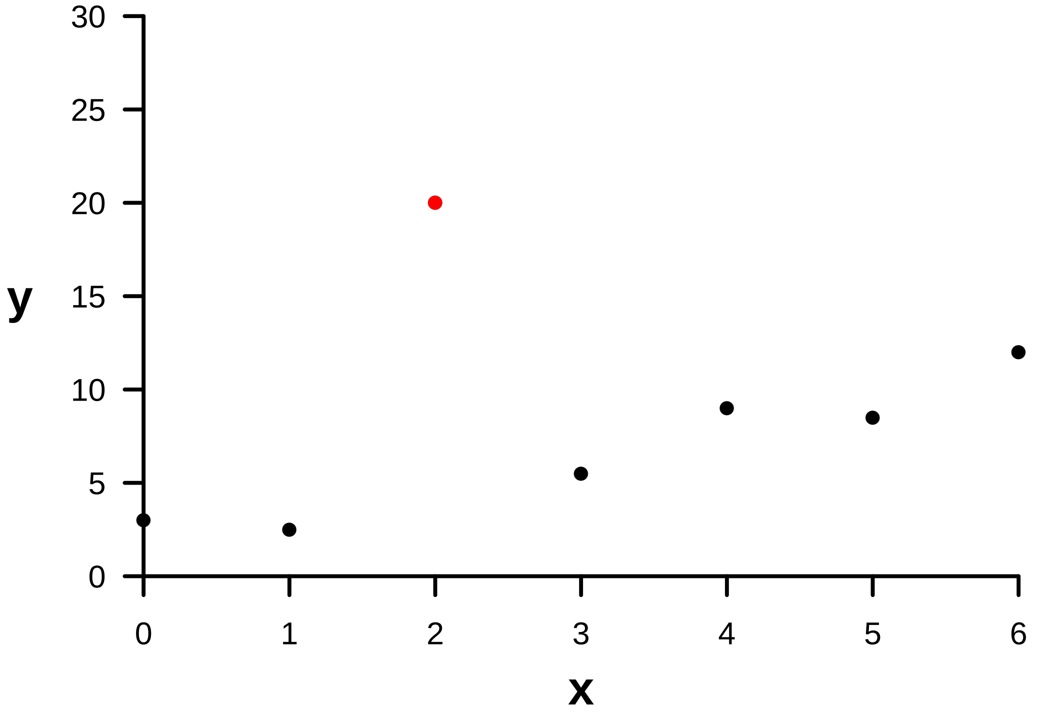
3.3. Worked Example

TIME TO FIT A MODEL

3.4. Linear models in R

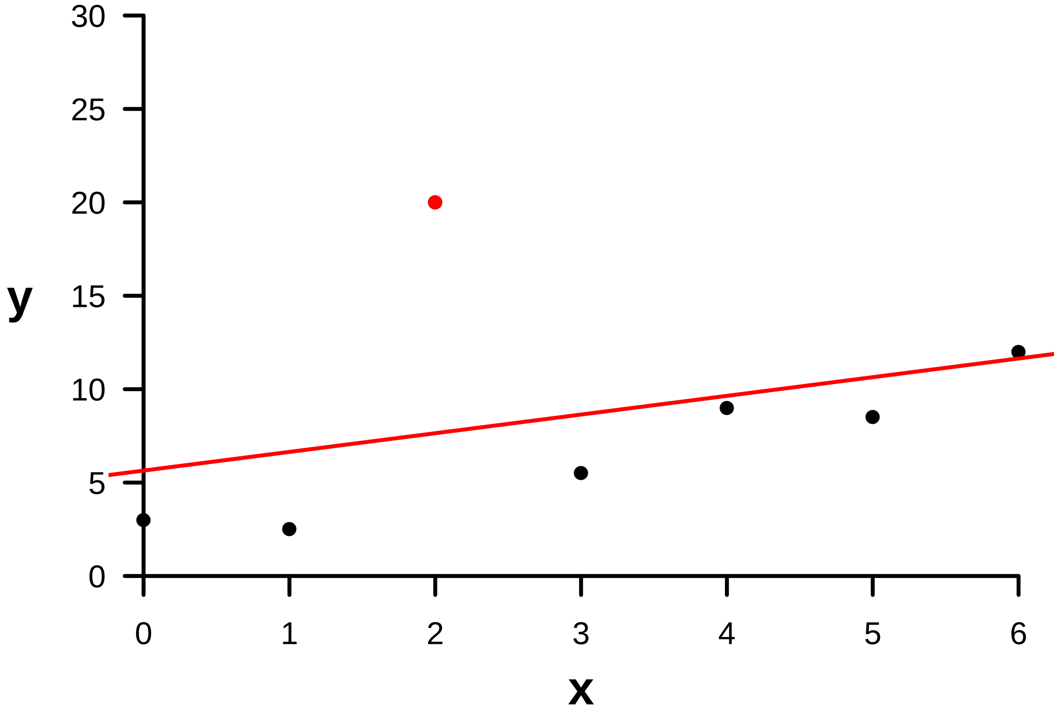
3.5. Model diagnostics

3.5.1. Residuals



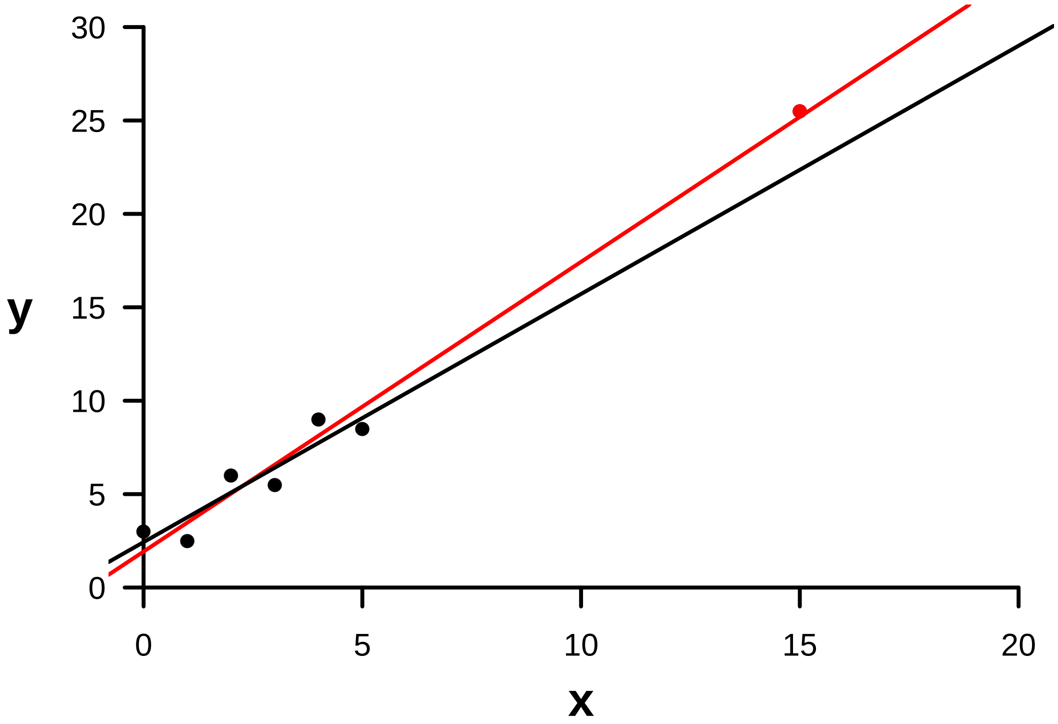
3.6. Model diagnostics

3.6.1. Residuals



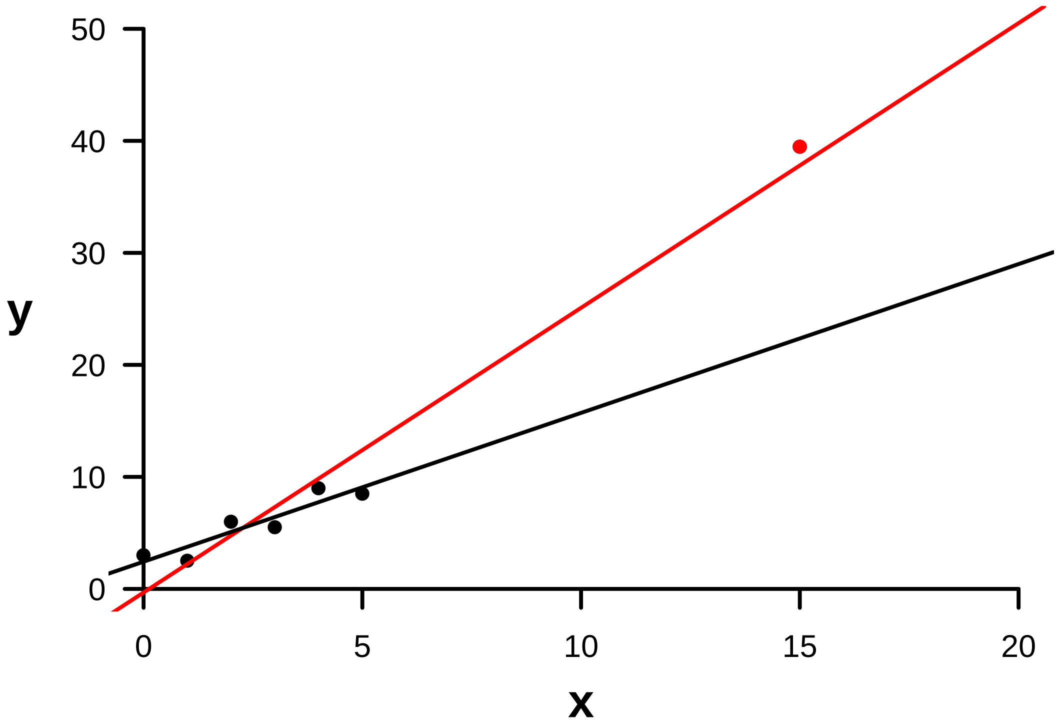
3.7. Model diagnostics

3.7.1. Leverage



3.8. Model diagnostics

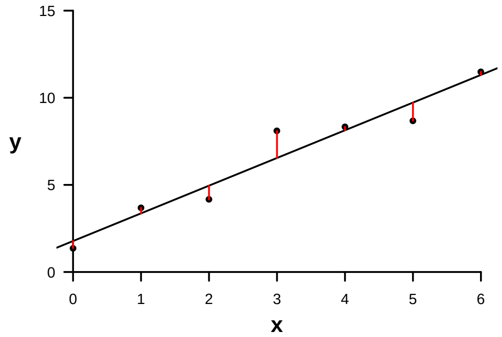
3.8.1. Cook's D



3.9. Example

3.9.1. Model evaluation

Extractor	Description
`residuals()`	Extracts residuals from model



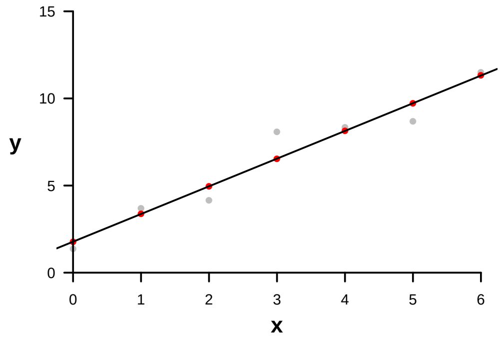

```
> residuals(DATA.lm)
```

```
        1          2          3          4          5          6          7 
0.8642857 -1.1428571  0.8500000 -1.1571429  0.8357143 -1.0714286  0.8214286 
```

3.10. Example

3.10.1. Model evaluation

Extractor	Description
`residuals()`	Extracts residuals from model
`fitted()`	Extracts the predicted values




```
> fitted(DATA.lm)
```

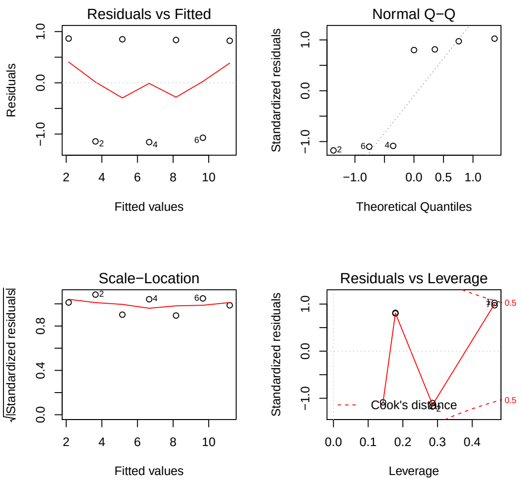
```
        1         2         3         4         5         6         7 
 2.135714  3.642857  5.150000  6.657143  8.164286  9.671429 11.178571 
```

3.11. Example

3.11.1. Model evaluation

Extractor	Description
`residuals()`	Extracts residuals from model
`fitted()`	Extracts the predicted values
`plot()`	Series of diagnostic plots

```
> plot(DATA.lm)
```



3.12. Example

3.12.1. Model evaluation

Extractor	Description
`residuals()`	Residuals
`fitted()`	Predicted values
`plot()`	Diagnostic plots
`influence.measures()`	Leverage (hat) and Cook's D

3.13. Example

3.13.1. Model evaluation

```
> influence.measures(DATA.lm)
```

```
Influence measures of
     lm(formula = Y ~ X, data = DATA) :

   dfb.1_     dfb.X  dffit cov.r cook.d   hat inf
1  0.9603 -7.99e-01  0.960  1.82 0.4553 0.464
2 -0.7650  5.52e-01 -0.780  1.15 0.2756 0.286
```

```
3  0.3165 -1.63e-01  0.365  1.43 0.0720 0.179
4 -0.2513 -7.39e-17 -0.453  1.07 0.0981 0.143
5  0.0443  1.60e-01  0.357  1.45 0.0696 0.179
6  0.1402 -5.06e-01 -0.715  1.26 0.2422 0.286
7 -0.3466  7.50e-01  0.901  1.91 0.4113 0.464
```

3.14. Example

3.14.1. Model evaluation

Extractor	Description
`residuals()`	Residuals
`fitted()`	Predicted values
`plot()`	Diagnostic plots
`influence.measures()`	Leverage, Cook's D
`summary()`	Summarizes important output from model

3.15. Example

3.15.1. Model evaluation

```
> summary(DATA.lm)
```

```
Call:
lm(formula = Y ~ X, data = DATA)

Residuals:
      1       2       3       4       5       6       7
 0.8643 -1.1429  0.8500 -1.1571  0.8357 -1.0714  0.8214

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1357     0.7850   2.721 0.041737 *
X             1.5071     0.2177   6.923 0.000965 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.152 on 5 degrees of freedom
Multiple R-squared:  0.9055,    Adjusted R-squared:  0.8866
F-statistic: 47.92 on 1 and 5 DF,  p-value: 0.0009648
```

3.16. Example

3.16.1. Model evaluation

Extractor	Description
`residuals()`	Residuals
`fitted()`	Predicted values

Extractor	Description
`plot()`	Diagnostic plots
`influence.measures()`	Leverage, Cook's D
`summary()`	Model output
`confint()`	Confidence intervals of parameters

3.17. Example

3.17.1. Model evaluation

```
> confint(DATA.lm)
```

```
                2.5 %   97.5 %
(Intercept) 0.1178919 4.153537
X           0.9474996 2.066786
```

3.18. Example

3.18.1. Model evaluation

Extractor	Description
`residuals()`	Residuals
`fitted()`	Predicted values
`plot()`	Diagnostic plots
`influence.measures()`	Leverage, Cook's D
`summary()`	Model output
`confint()`	Confidence intervals
`predict()`	Predict responses to new levels of predictors

3.19. Example

3.19.1. Model evaluation

```
> predict(DATA.lm, newdata=data.frame(X=c(2.5, 4.1)),
+         se=TRUE)
```

```
$fit
       1        2
5.903571 8.315000

$se.fit
        1         2
0.4488222 0.4969340

$df
[1] 5
```

```
$residual.scale
[1] 1.152017
```

```
> predict(DATA.lm, newdata=data.frame(X=c(2.5, 4.1)),
+         interval='confidence')
```

```
       fit      lwr      upr
1 5.903571 4.749837 7.057306
2 8.315000 7.037591 9.592409
```

3.20. Example

3.20.1. Model evaluation

```
> predict(DATA.lm, newdata=data.frame(X=c(2.5, 4.1)),
+         interval='prediction')
```

```
       fit      lwr       upr
1 5.903571 2.725409  9.081734
2 8.315000 5.089881 11.540119
```

3.21. Prediction

$$\underbrace{\begin{pmatrix}3.0\\2.5\\6.0\\5.5\\9.0\\8.6\\12.0\end{pmatrix}}_{\text{Response values}} = \underbrace{\begin{pmatrix}1&0\\1&1\\1&2\\1&3\\1&4\\1&5\\1&6\end{pmatrix}}_{\text{Model matrix}} \times \underbrace{\begin{pmatrix}\beta_0\\\beta_1\end{pmatrix}}_{\text{Parameter vector}} + \underbrace{\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\\\varepsilon_6\end{pmatrix}}_{\text{Residual vector}}$$

$$\underbrace{\begin{pmatrix}1&0\\1&1\\1&2\\1&3\\1&4\\1&5\\1&6\end{pmatrix}}_{\text{Model matrix}} \times \underbrace{\begin{pmatrix}2.136\\1.507\end{pmatrix}}_{\text{Parameter vector}} = \underbrace{\begin{pmatrix}2.136\\3.643\\5.150\\6.657\\8.164\\9.671\\11.179\end{pmatrix}}_{\text{Predicted values vector}}$$

3.22. Example

3.22.1. Model evaluation

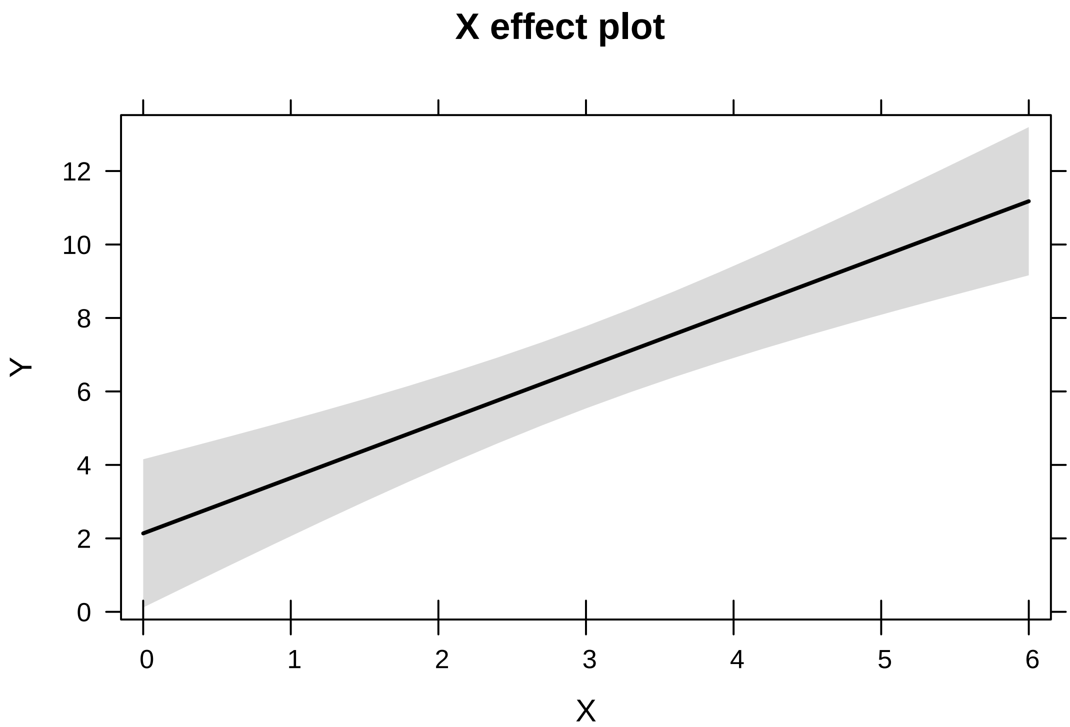
Extractor	Description
`residuals()`	Residuals
`fitted()`	Predicted values
`plot()`	Diagnostic plots
`influence.measures()`	Leverage, Cook's D

Extractor	Description
`summary()`	Model output
`confint()`	Confidence intervals
`predict()`	Predict new responses
`plot(allEffects())`	Effects plots

3.23. Example

3.23.1. Model evaluation

```
> library(effects)
> plot(allEffects(DATA.lm))
```



4. Worked Examples

4.1. Worked Examples

```
> fert <- read.csv('../data/fertilizer.csv', strip.white=T)
> fert
```

```
  FERTILIZER YIELD
1         25    84
2         50    80
3         75    90
4        100   154
5        125   148
6        150   169
7        175   206
```

```
8          200   244
9          225   212
10         250   248
```

```
> head(fert)
```

```
  FERTILIZER YIELD
1         25    84
2         50    80
3         75    90
4        100   154
5        125   148
6        150   169
```

```
> summary(fert)
```

```
   FERTILIZER         YIELD
 Min.   : 25.00   Min.   : 80.0
 1st Qu.: 81.25   1st Qu.:104.5
 Median :137.50   Median :161.5
 Mean   :137.50   Mean   :163.5
 3rd Qu.:193.75   3rd Qu.:210.5
 Max.   :250.00   Max.   :248.0
```

```
> str(fert)
```

```
'data.frame':   10 obs. of  2 variables:
 $ FERTILIZER: int  25 50 75 100 125 150 175 200 225 250
 $ YIELD     : int  84 80 90 154 148 169 206 244 212 248
```

4.2. Worked Examples

Question: is there a relationship between fertilizer concentration and grass yeild?

Linear model:

$$Y * i = \beta * 0 + \beta_1 F_i + \varepsilon_i \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

4.3. Worked Examples

```
> peake <- read.csv('../data/peakquinn.csv', strip.white=T)
> head(peake)
```

```
     AREA INDIV
1  516.00    18
2  469.06    60
3  462.25    57
4  938.60   100
5 1357.15    48
6 1773.66   118
```

```
> summary(peake)
```

```
      AREA              INDIV
 Min.   :  462.2   Min.   :  18.0
 1st Qu.: 1773.7   1st Qu.: 148.0
 Median : 4451.7   Median : 338.0
 Mean   : 7802.0   Mean   : 446.9
 3rd Qu.: 9287.7   3rd Qu.: 632.0
 Max.   :27144.0   Max.   :1402.0
```

4.4. Worked Examples

Question: is there a relationship between mussel clump area and number of individuals?

Linear model:

$$Indiv_i = \beta_0 + \beta_1 Area_i + \varepsilon_i \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$
$$ln(Indiv_i) = \beta_0 + \beta_1 ln(Area_i) + \varepsilon_i \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$