Workshop 7.2a: Introduction to Linear models

Murray Logan

19 Jul 2017

Section 1

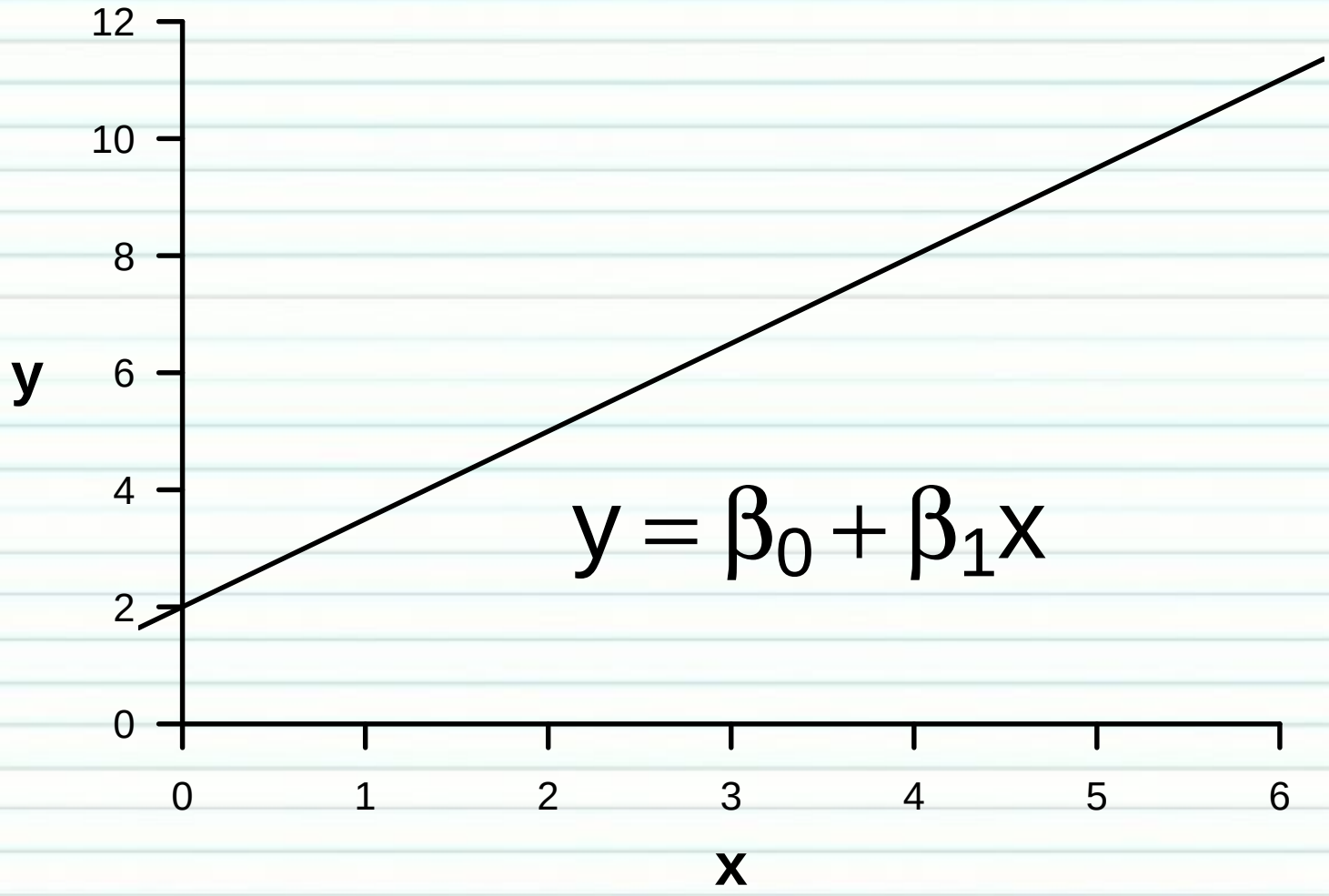
Revision

Aims of statistical modelling

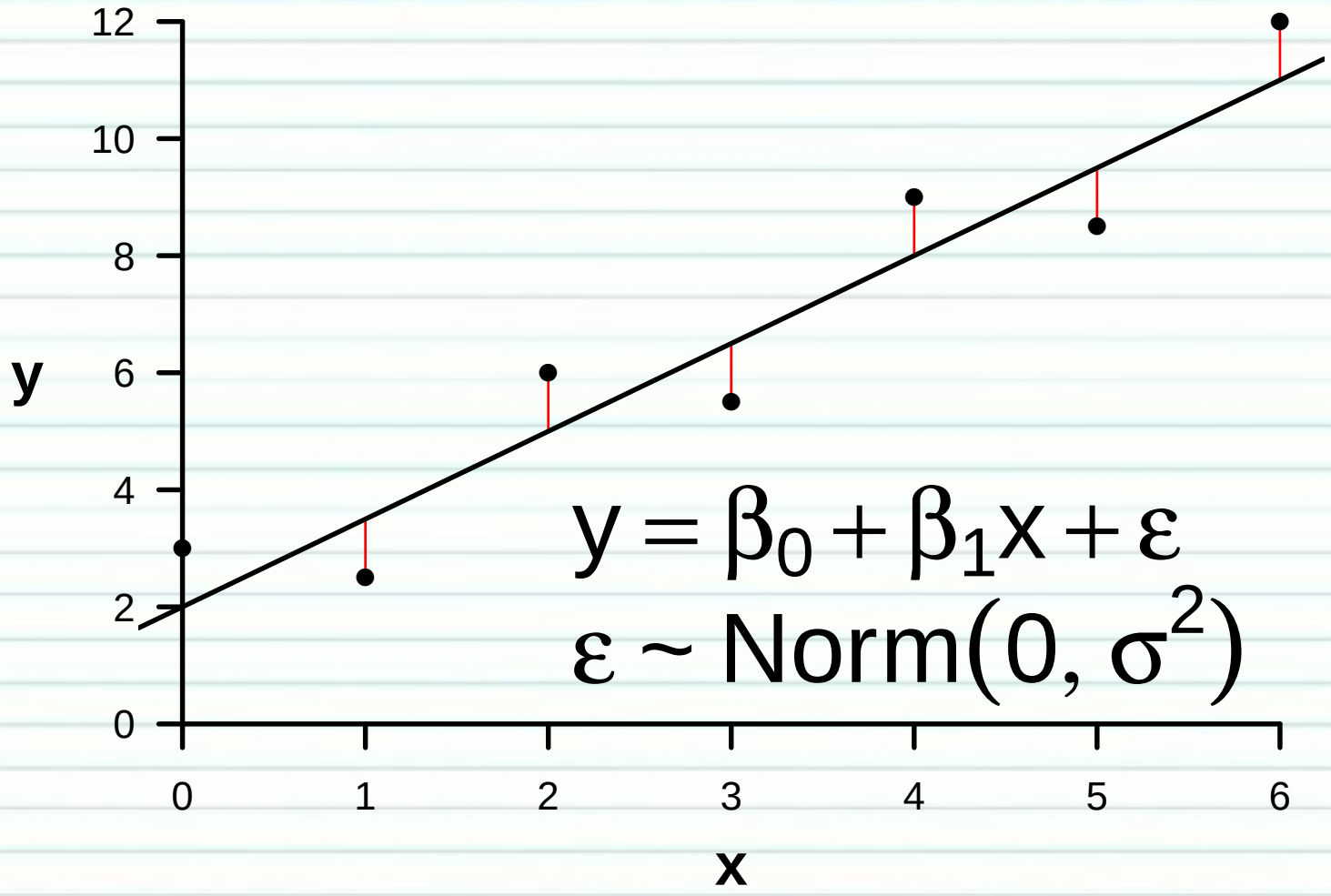
Use samples to:

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

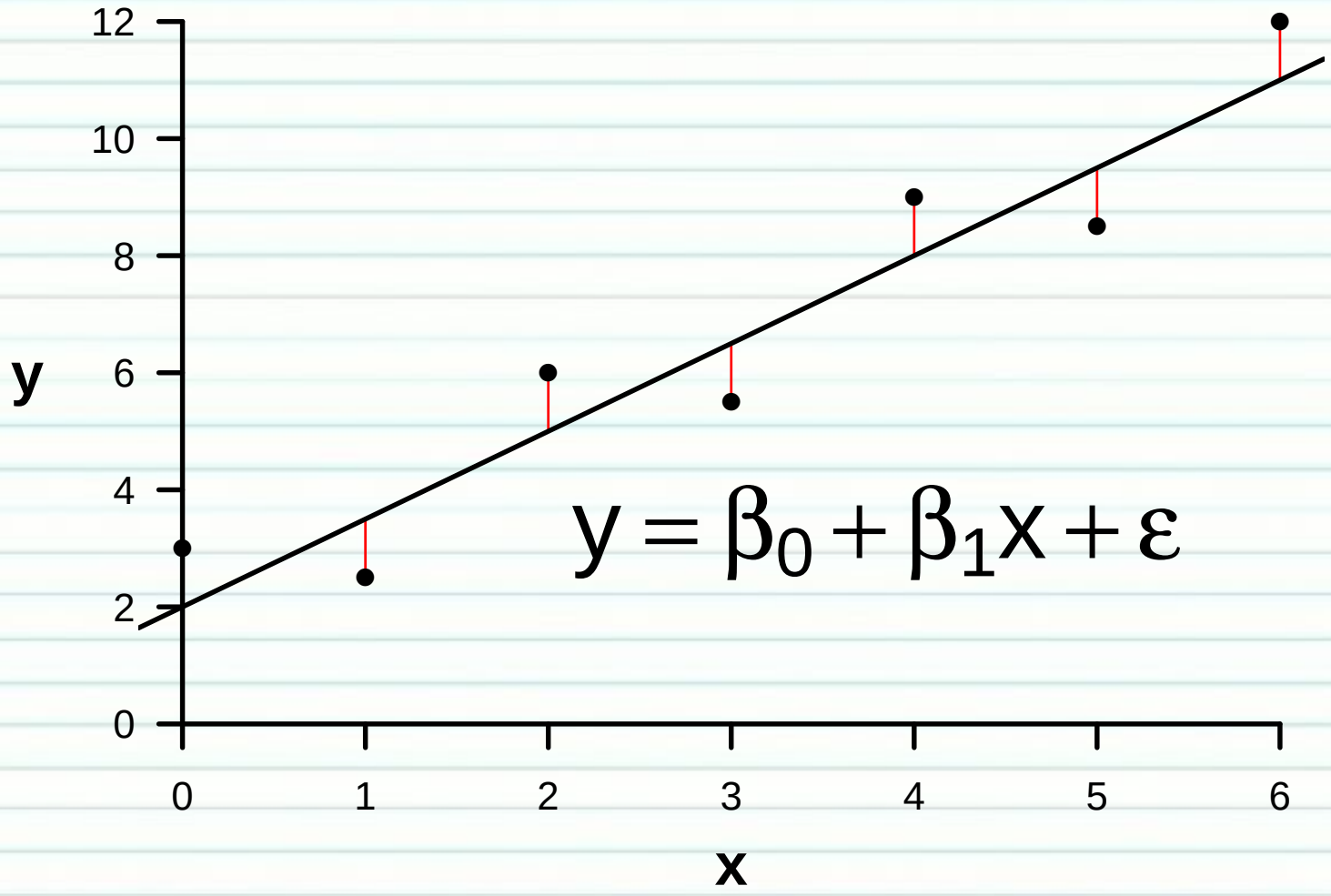
Mathematical models



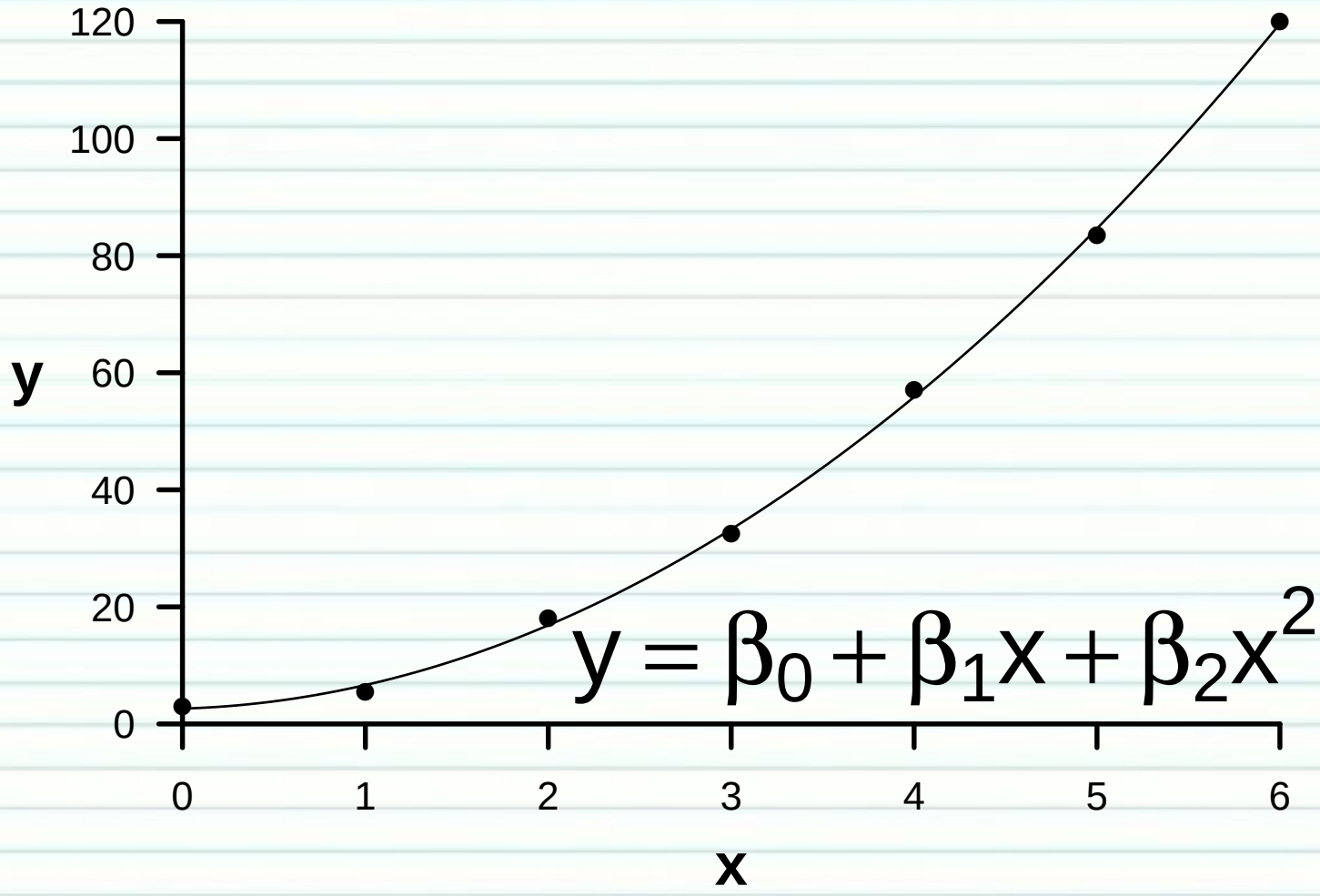
Statistical models



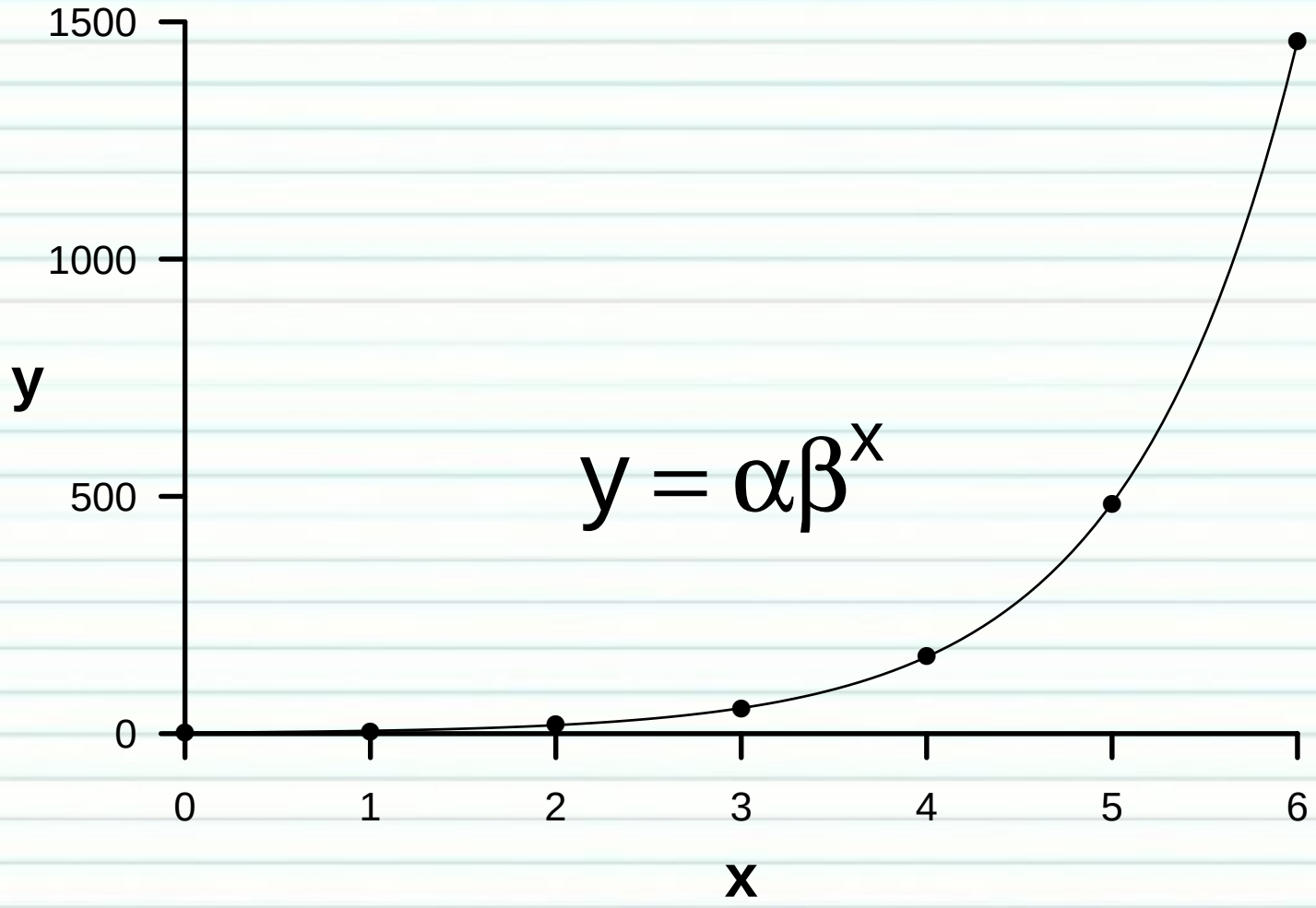
Linear models



Linear models



Non-linear models



Linear models

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

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

Linear models

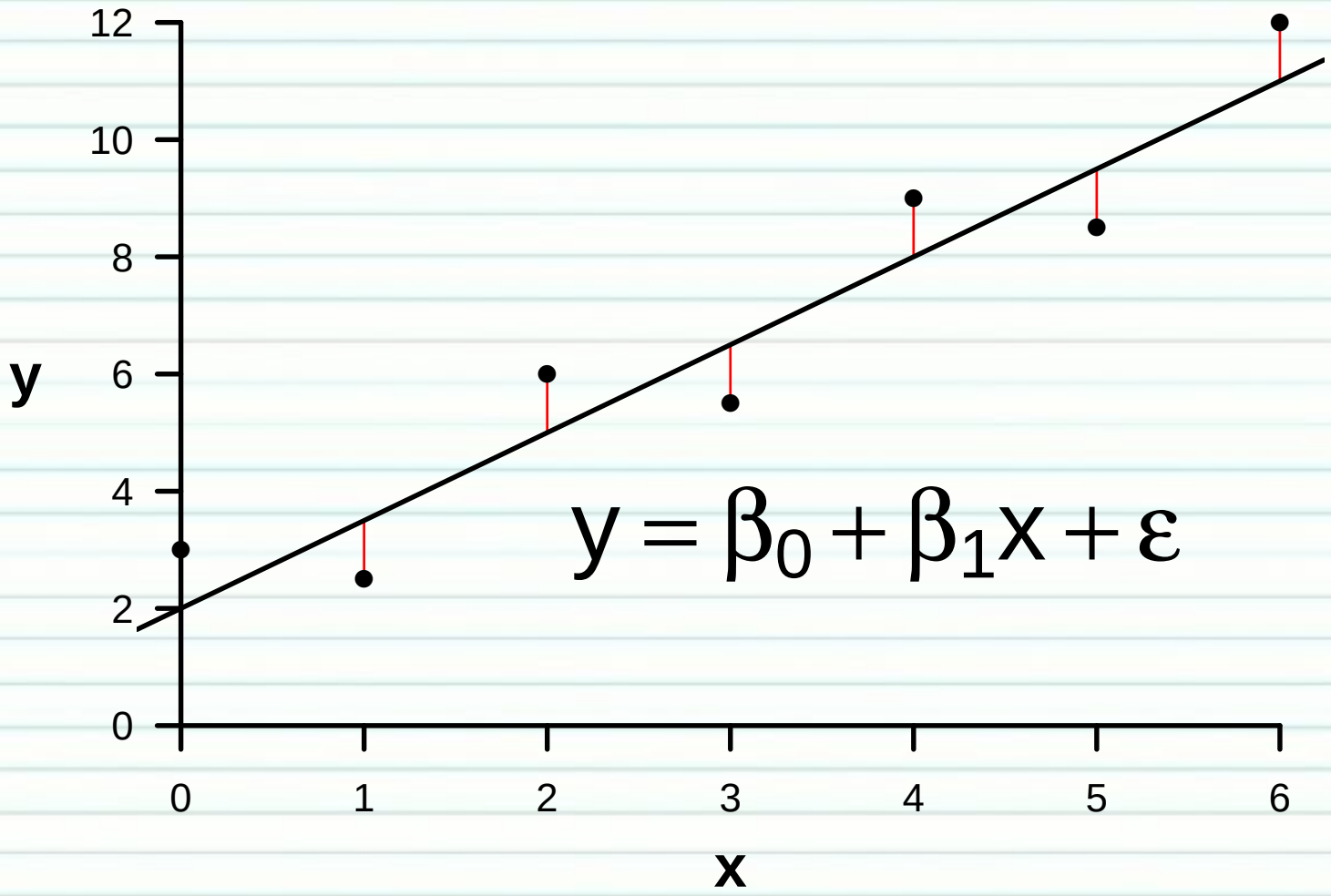
$$\mathbf{y}_i = \beta_0 + \beta_1 \times \mathbf{x}_1 + \varepsilon_1$$

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

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

Estimation



Ordinary Least Squares

Estimation

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

$$
\begin{aligned}
3.0 &= \beta_0 \times 1 + \beta_1 \times 0 + \varepsilon_1 \\
2.5 &= \beta_0 \times 1 + \beta_1 \times 1 + \varepsilon_1 \\
6.0 &= \beta_0 \times 1 + \beta_1 \times 2 + \varepsilon_2 \\
5.5 &= \beta_0 \times 1 + \beta_1 \times 3 + \varepsilon_3
\end{aligned}
$$

Estimation

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

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

Inference testing

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

The t-statistic

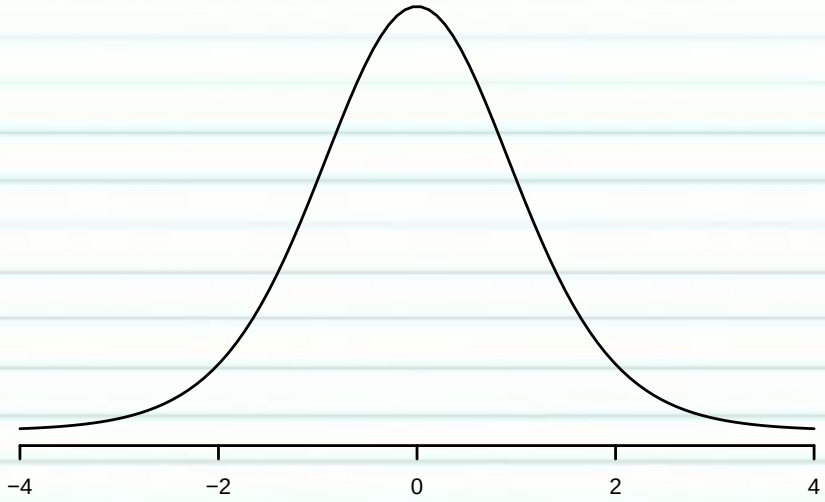
$$t = \frac{\text{param}}{SE_{\text{param}}}$$

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

Inference testing

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

The t-statistic and the t distribution



Section 2

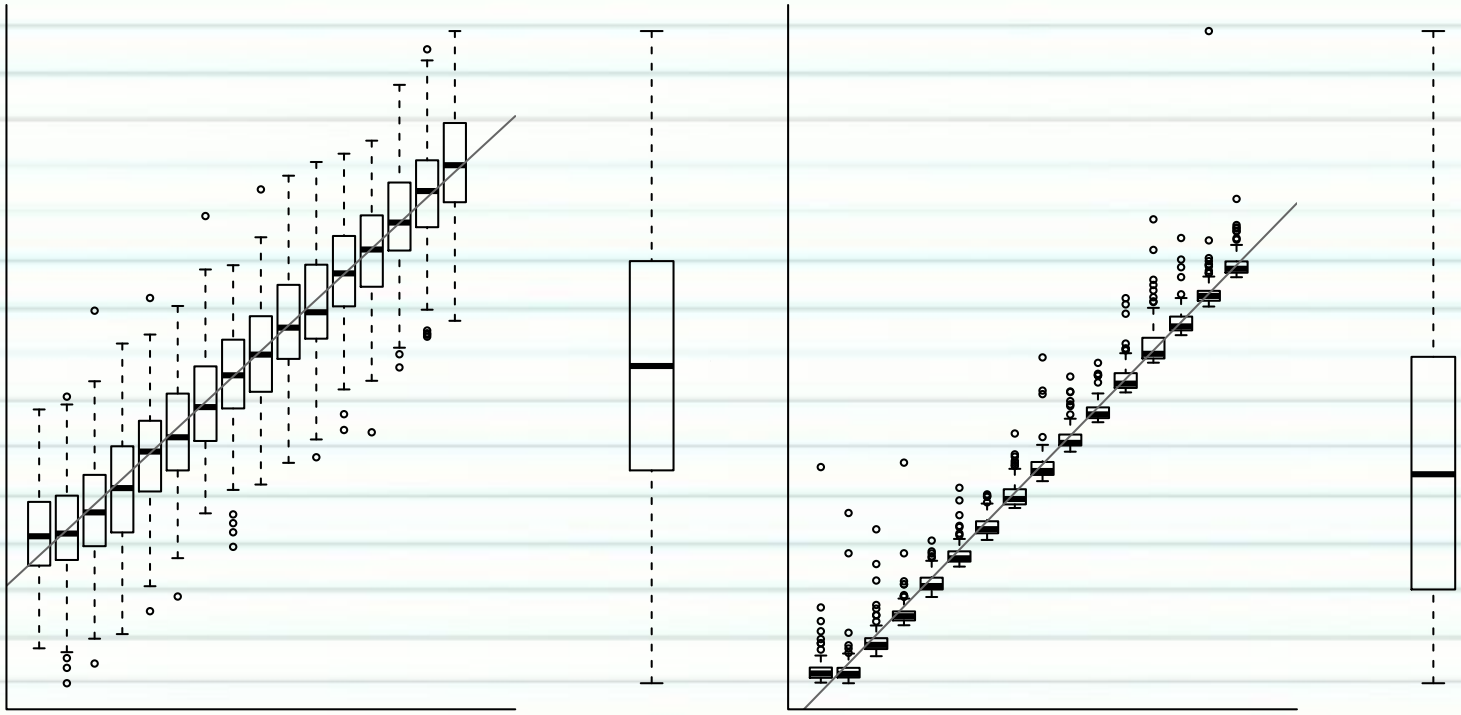
Linear model Assumptions

Assumptions

- Independence - unbiased, scale of treatment
- Normality - residuals
- Homogeneity of variance - residuals
- Linearity

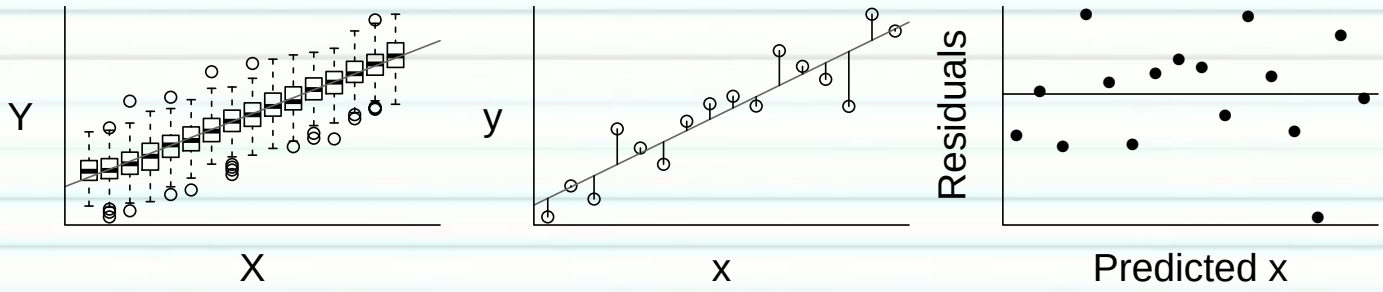
Assumptions

NORMALITY



Assumptions

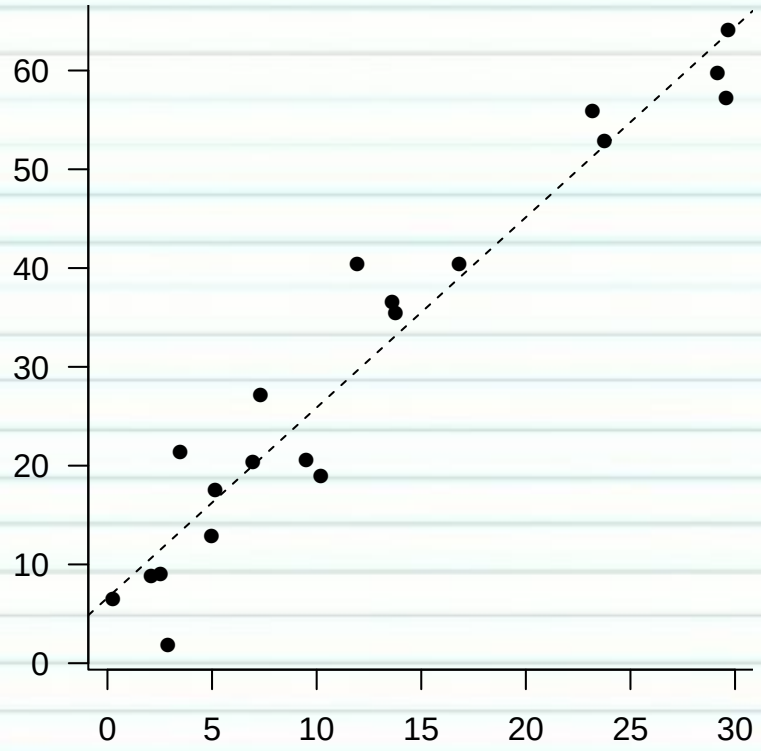
HOMOGENEITY OF VARIANCE

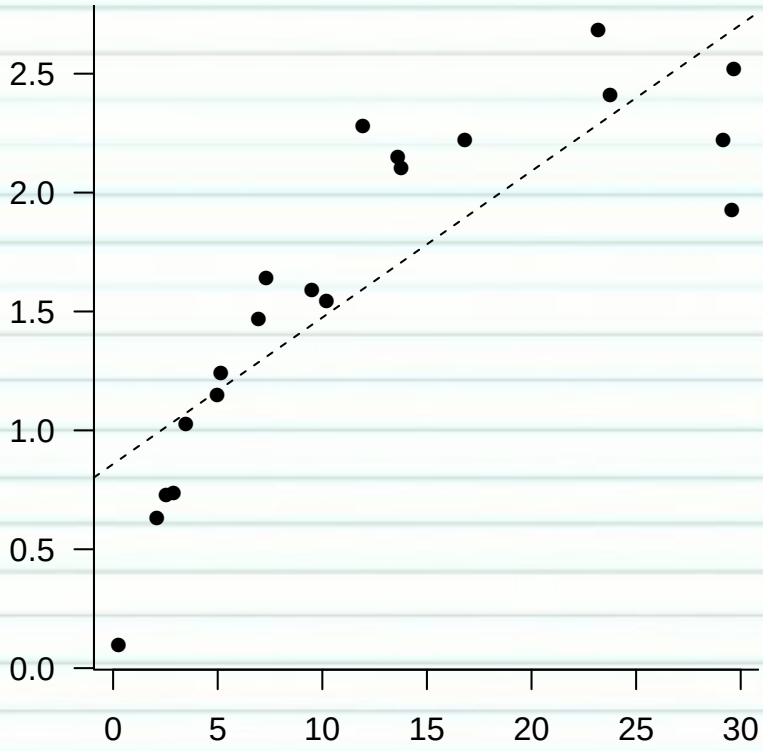


Assumptions

LINEARITY

Trendline

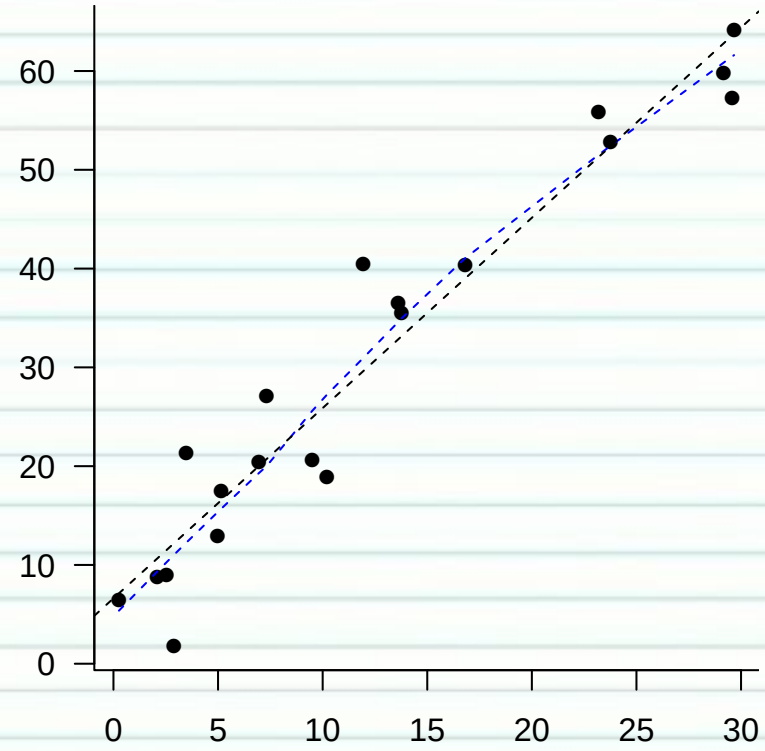




Assumptions

LINEARITY

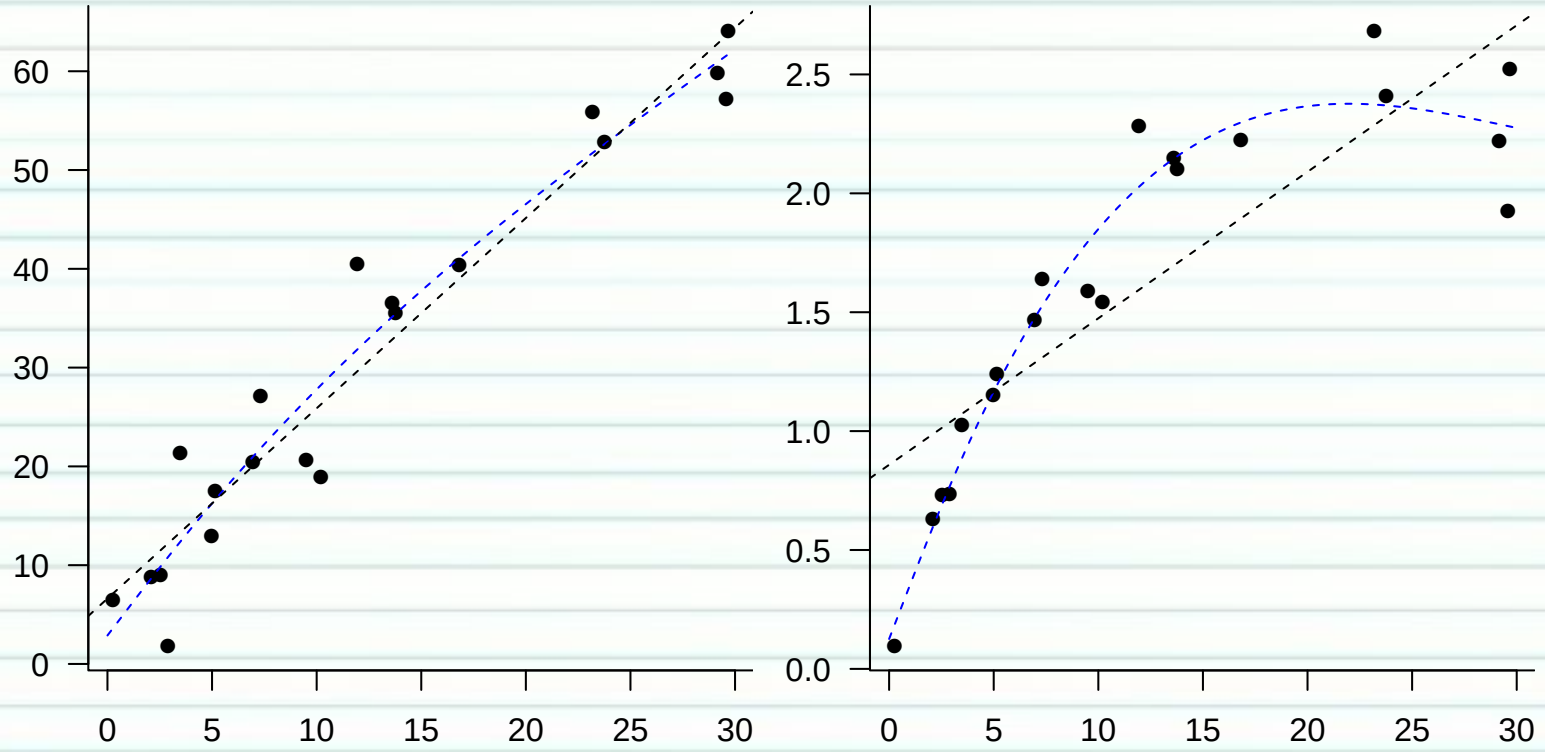
Loess (lowess) smoother



Assumptions

LINEARITY

Spline smoother



Assumptions

$$\mathtt{y_i} = \beta_0 + \beta_1 \times \mathtt{x_i} + \varepsilon_\mathtt{i}$$

$$\epsilon_\mathtt{i} \sim \mathcal{N}(0, \sigma^2)$$

Assumptions

$$\mathtt{y_i} = \beta_0 + \beta_1 \times \mathtt{x_i} + \varepsilon_\mathtt{i}$$

$$\epsilon_\mathtt{i} \sim \mathcal{N}(0, \sigma^2)$$

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

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)
```

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
```

Worked Examples

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

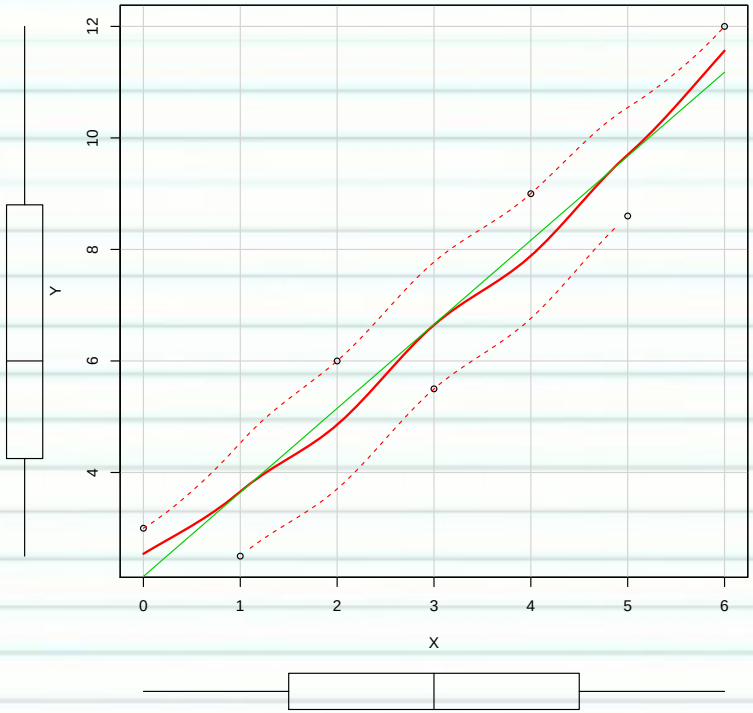
Linear model:

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

Example

EXPLORATORY DATA ANALYSIS

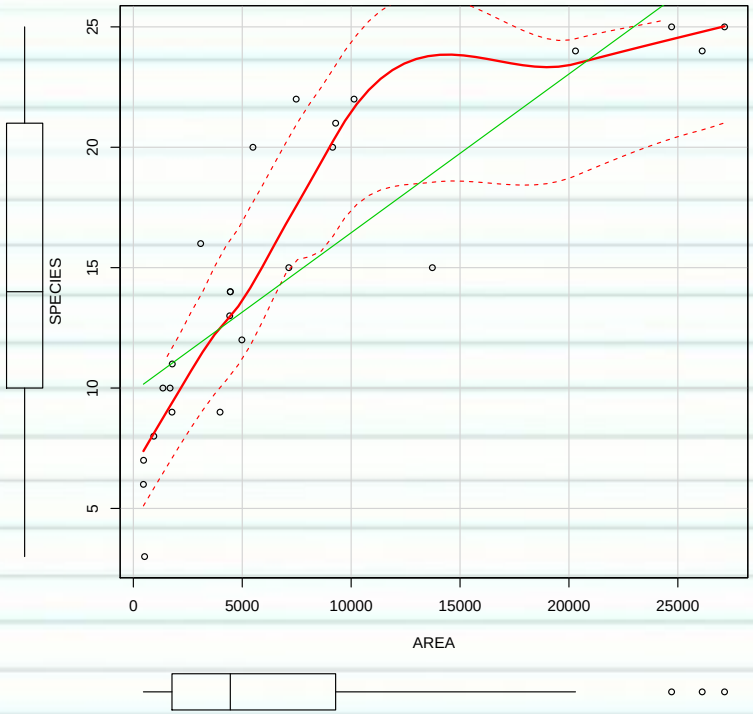
```
> library(car)
> scatterplot(Y~X, data=DATA)
```



Example

EXPLORATORY DATA ANALYSIS

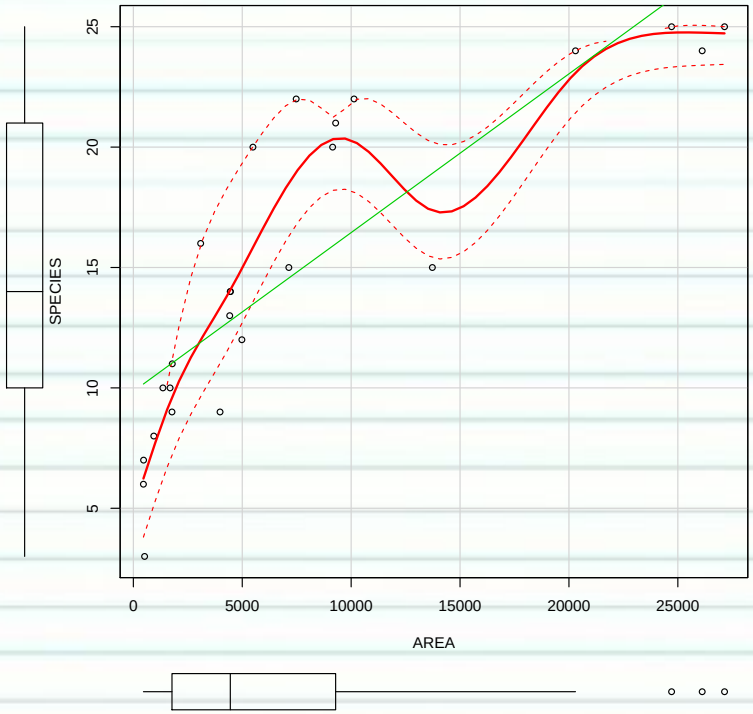
```
> library(car)
> peake <- read.csv('../data/peake.csv')
> scatterplot(SPECIES ~ AREA, data=peake)
```



Example

EXPLORATORY DATA ANALYSIS

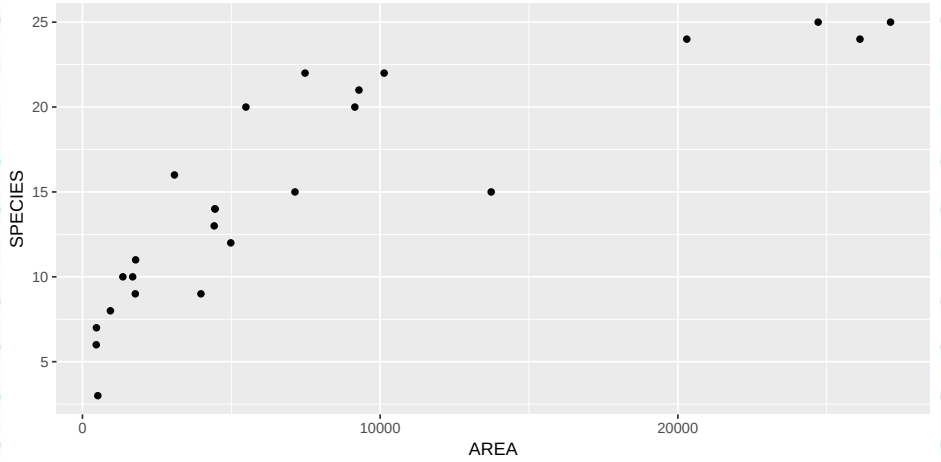
```
> scatterplot(SPECIES ~ AREA, data=peake,
+                 smoother=gamLine)
```



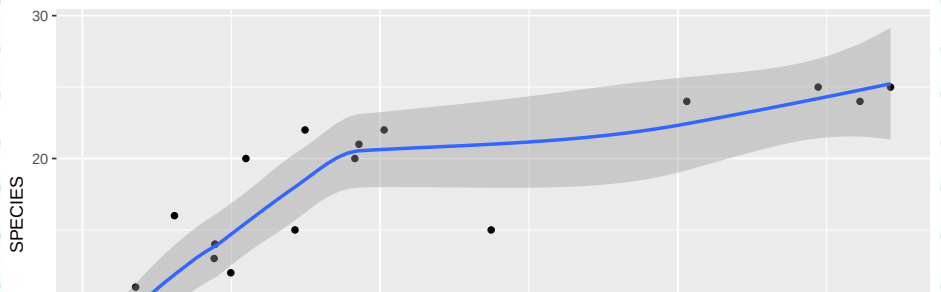
Example

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()
```



Section 3

Simple Linear models in R

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

Example

FIT LINEAR MODEL

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

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

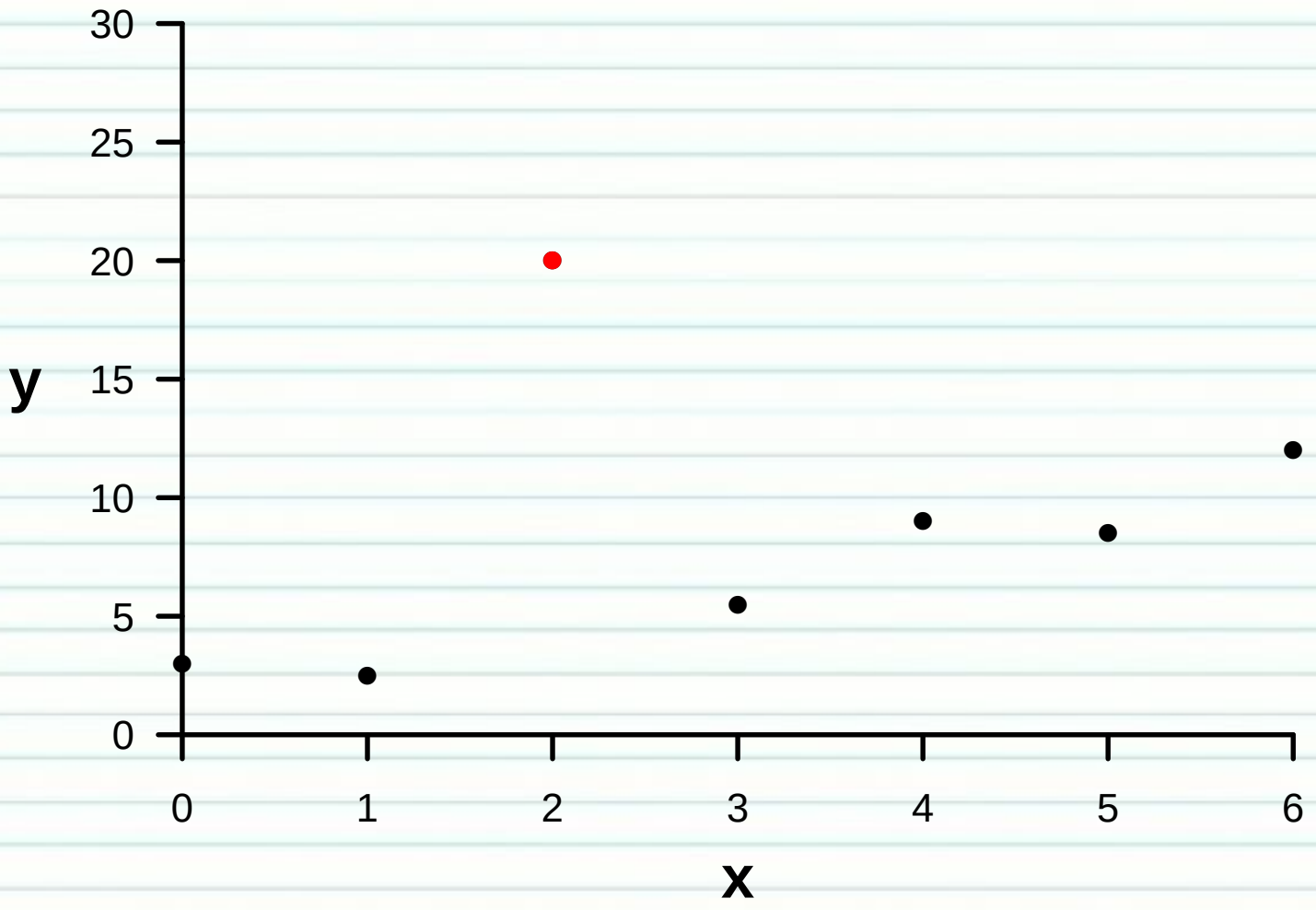
Worked Example

TIME TO FIT A MODEL

Linear models in R

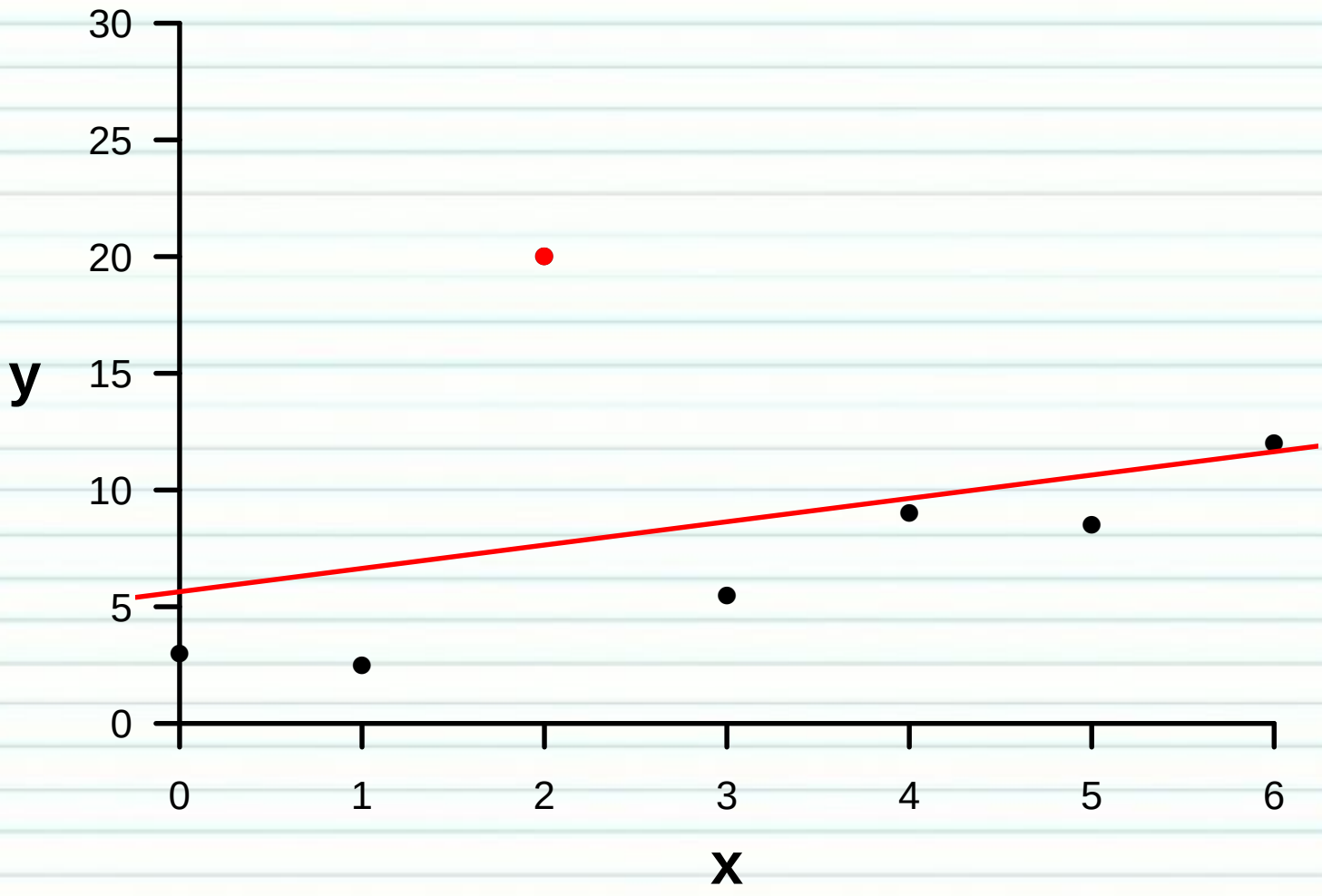
Model diagnostics

RESIDUALS



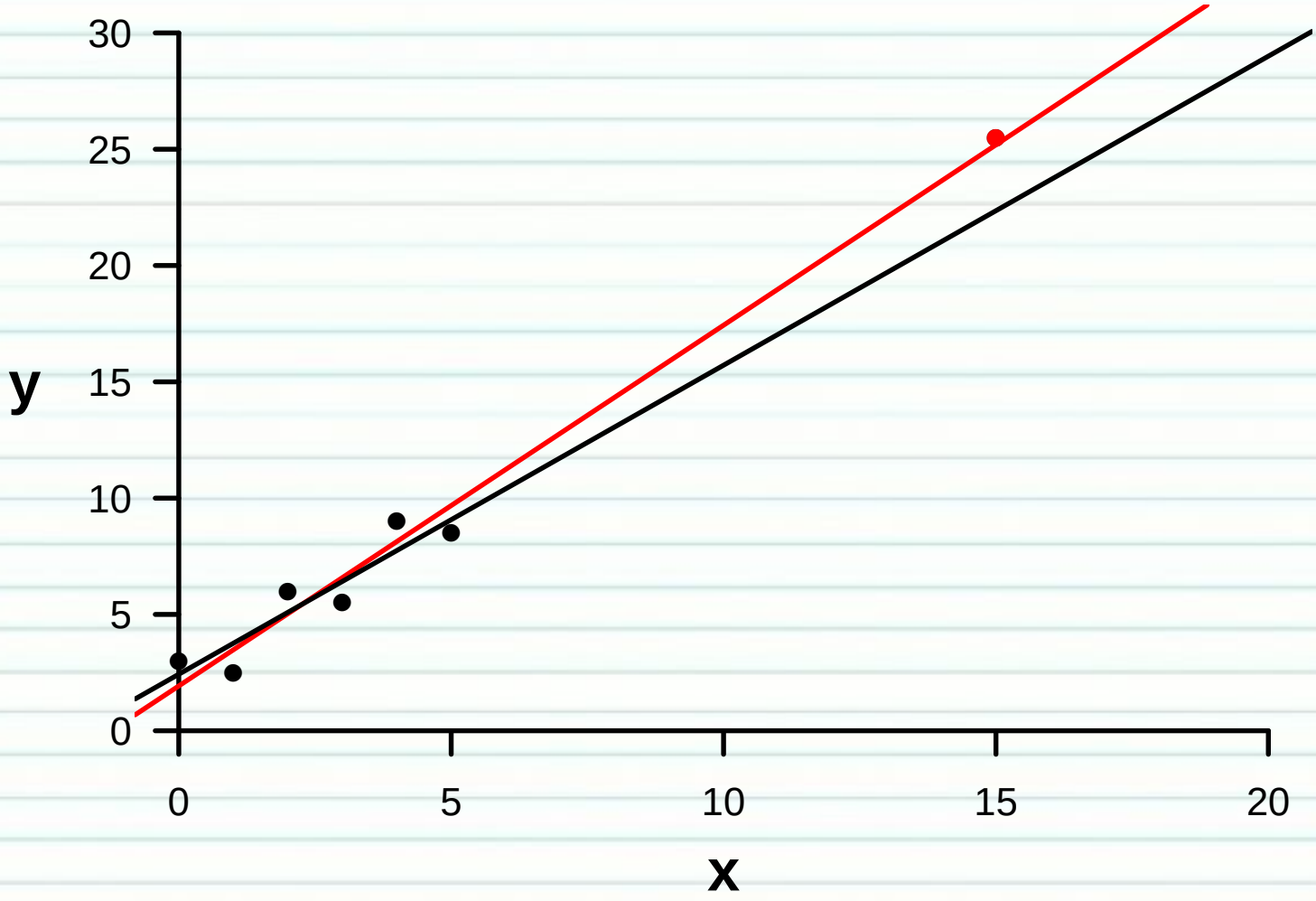
Model diagnostics

RESIDUALS



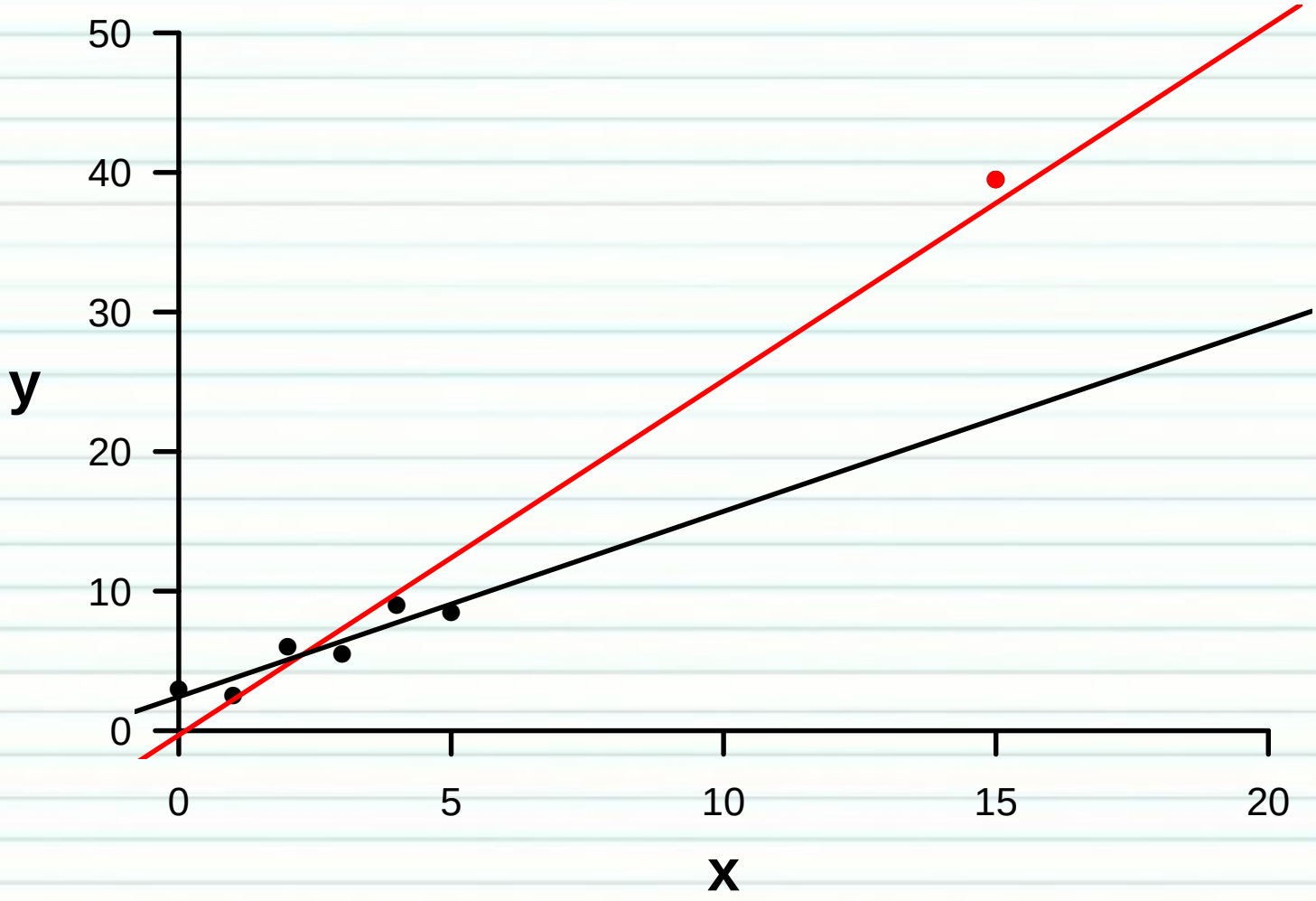
Model diagnostics

LEVERAGE



Model diagnostics

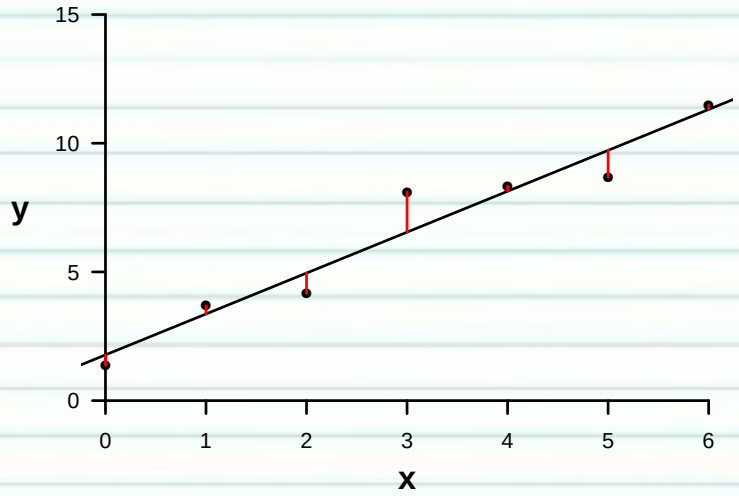
COOK'S D



Example

MODEL EVALUATION

Extractor	Description
residuals()	Extracts residuals from model

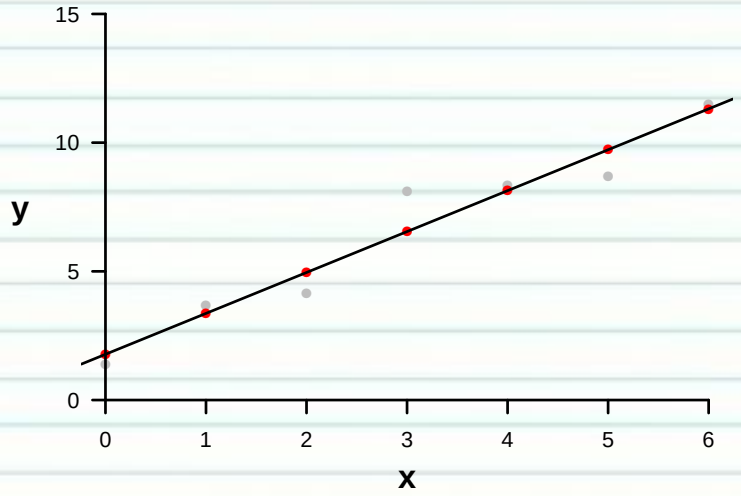



```
> residuals(DATA.lm)
```

Example

MODEL EVALUATION

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



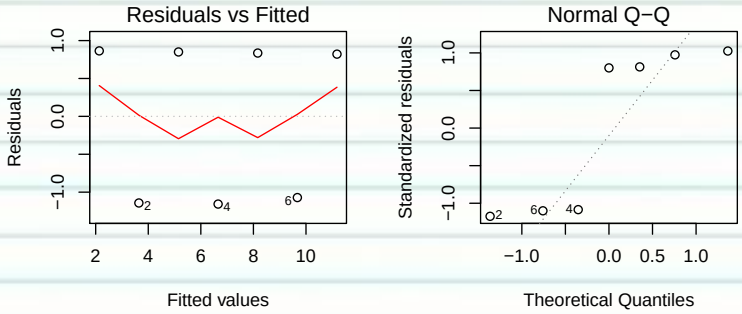

```
> fitted(DATA.lm)
```

Example

MODEL EVALUATION

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

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



Example

MODEL EVALUATION

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

Example

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
```

Example

MODEL EVALUATION

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

Example

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
```

Example

MODEL EVALUATION

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

Example

MODEL EVALUATION

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

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

Example

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 predic

Example

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')
```

Example

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
```

Prediction

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

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

Example

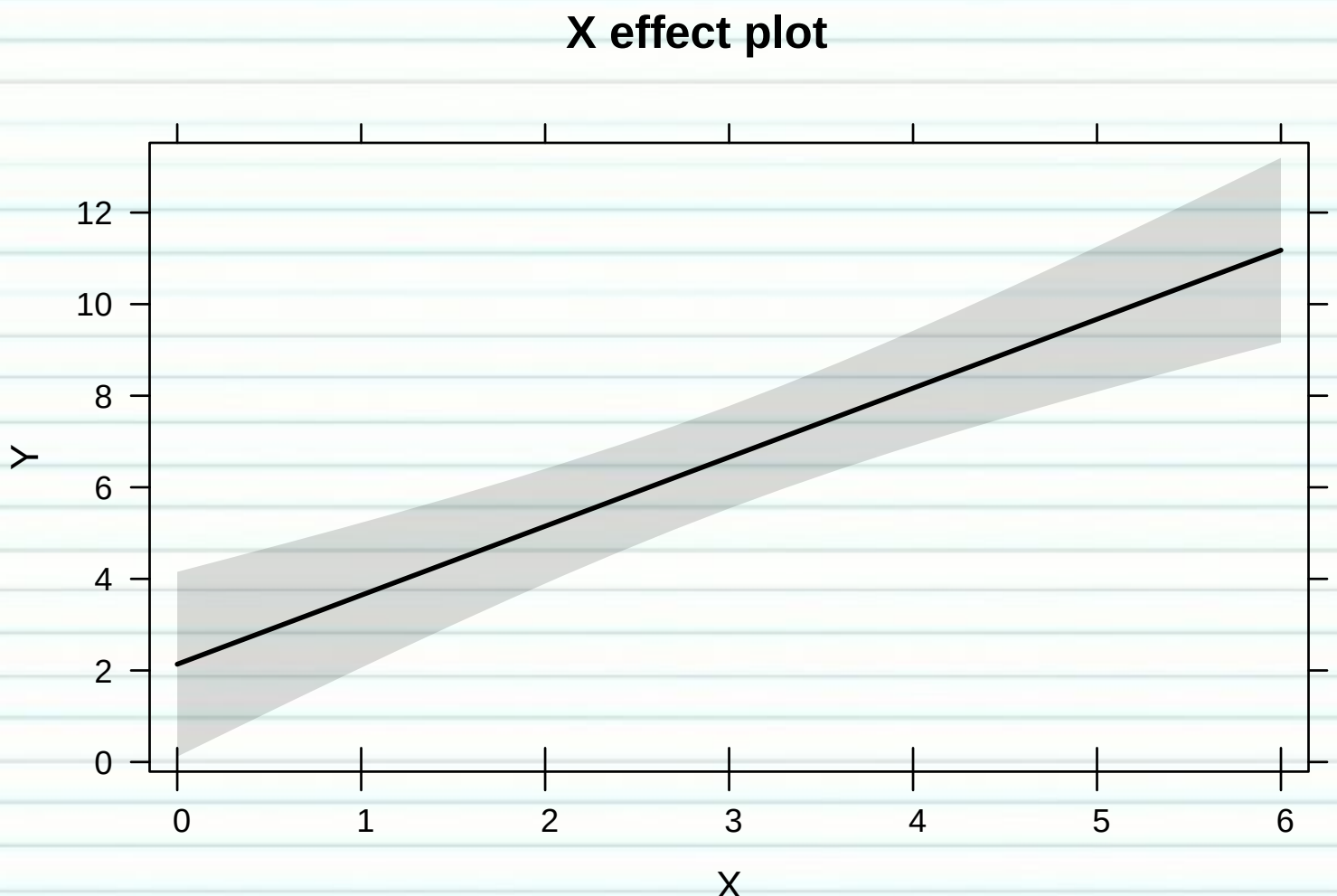
MODEL EVALUATION

Extractor	Description
`residuals()`	Residuals
`fitted()`	Predicted values
`plot()`	Diagnostic plots
`influence.measures()`	Leverage, Cook□s D
`summary()`	Model output
`confint()`	Confidence intervals
`predict()`	Predict new responses
`plot(allEffects())`	Effects plots

Example

MODEL EVALUATION

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



Section 4

Worked Examples

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
```

Worked Examples

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

Linear model:

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

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
```

Worked Examples

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

Linear model:

$$\text{Indiv}_i = \beta_0 + \beta_1 \text{Area}_i + \varepsilon_i \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\ln(\text{Indiv}_i) = \beta_0 + \beta_1 \ln(\text{Area}_i) + \varepsilon_i \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$