Linear Modelling: Simple Regression

7th of February 2023

R. Nicholls / D.-L. Couturier / C.S.R. Chilamakuri

Introduction:

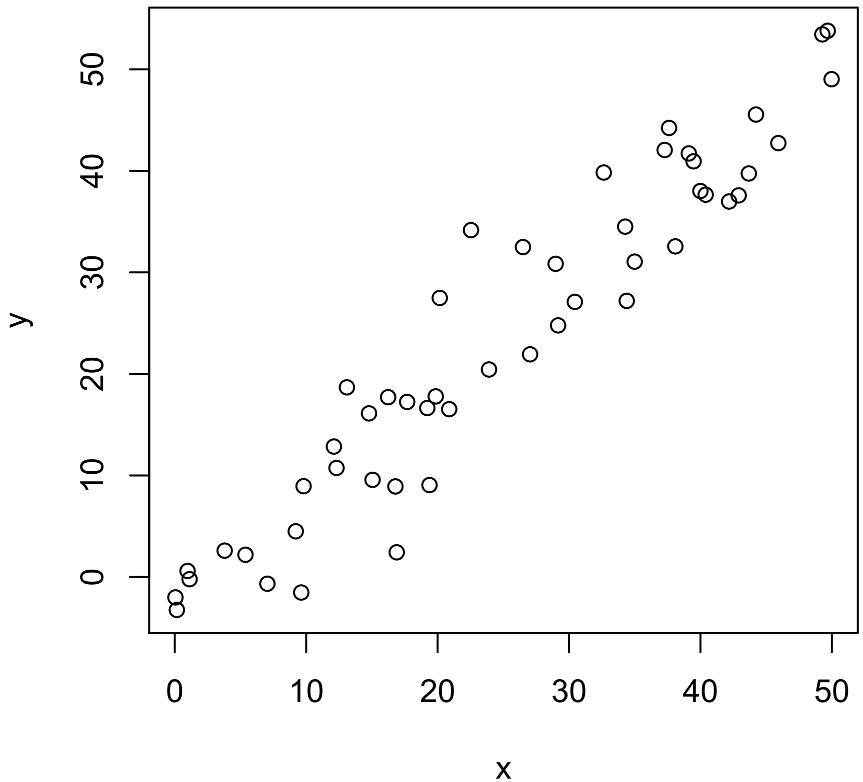
ANOVA

- Used for testing hypotheses regarding differences between groups
- Considers the variation within and between groups

Regression

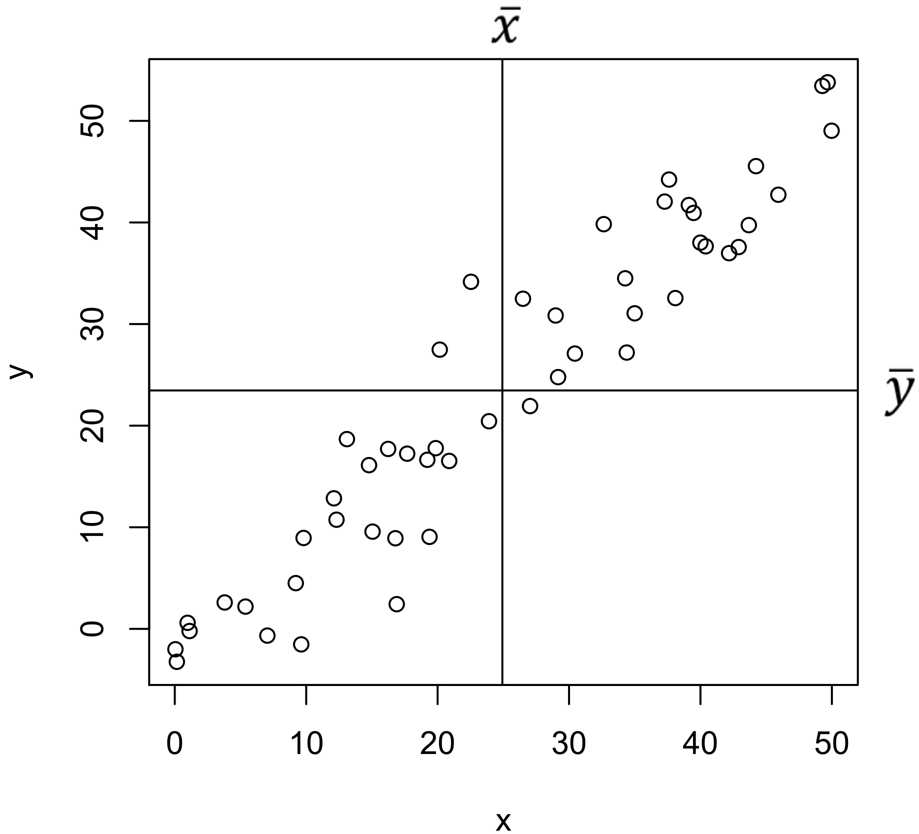
- Used for revealing and investigating relationships between input and output variables
- Aim is to model data, and extrapolate as much information as possible

Correlation:

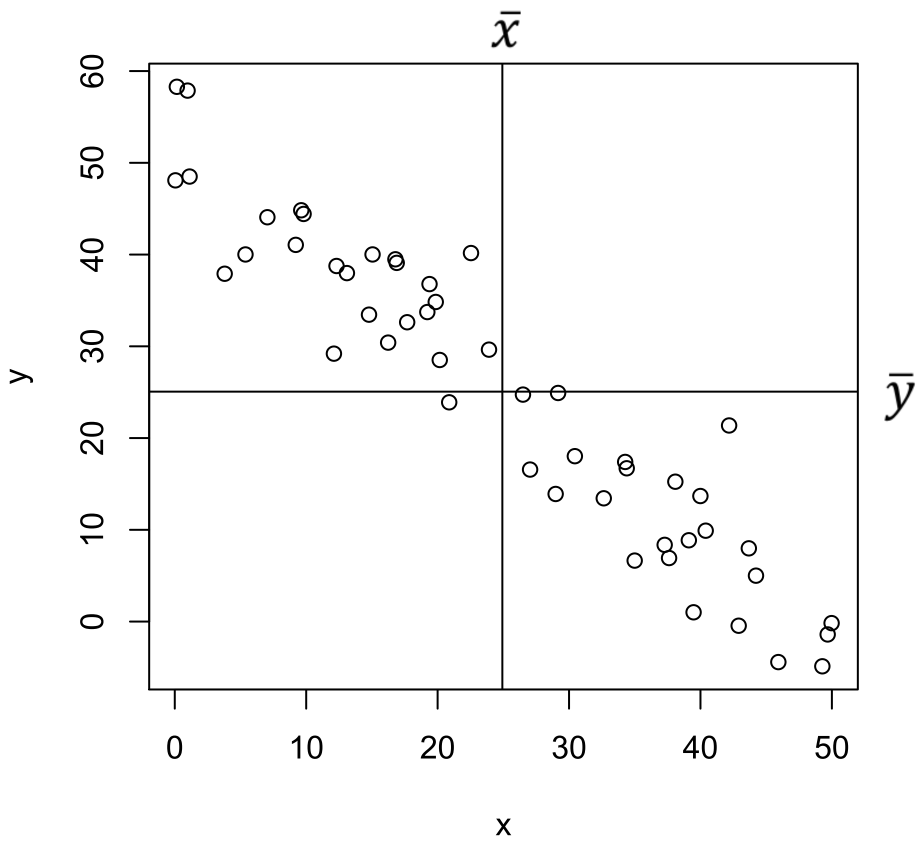


How to measure the strength of a linear relationship between variables?

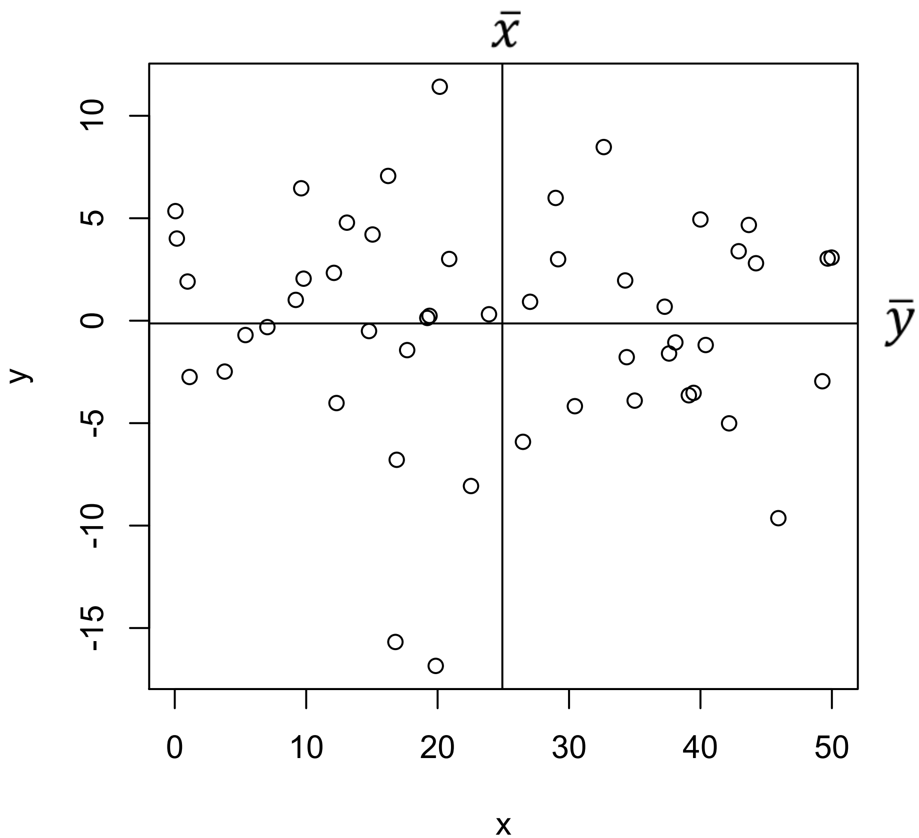
Correlation:



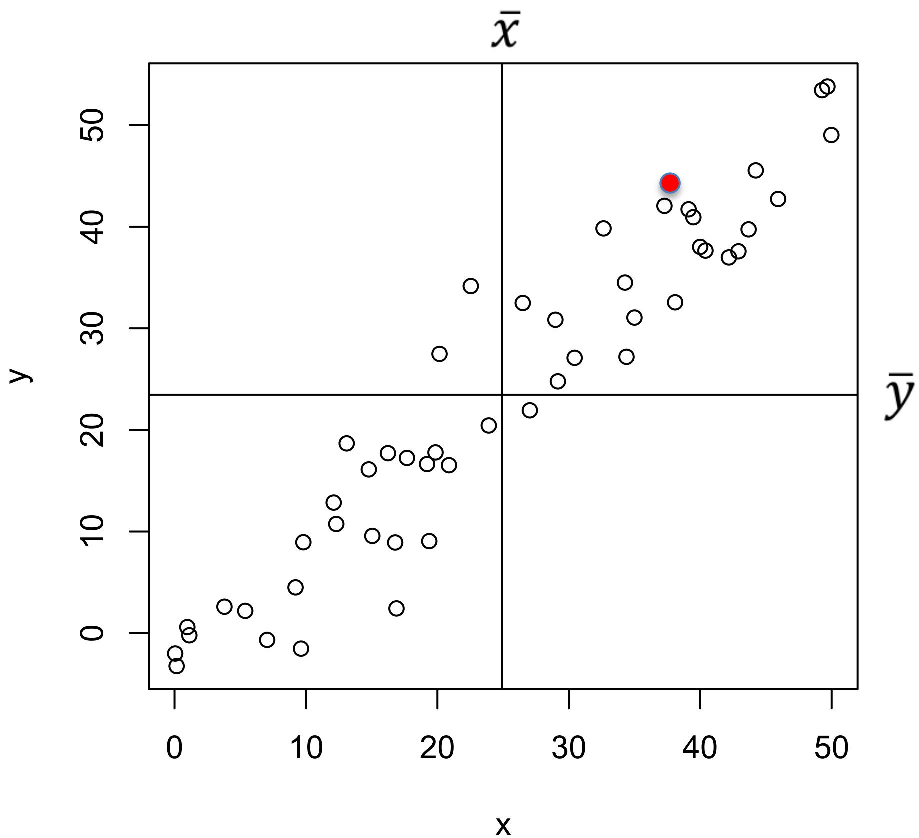
Correlation:



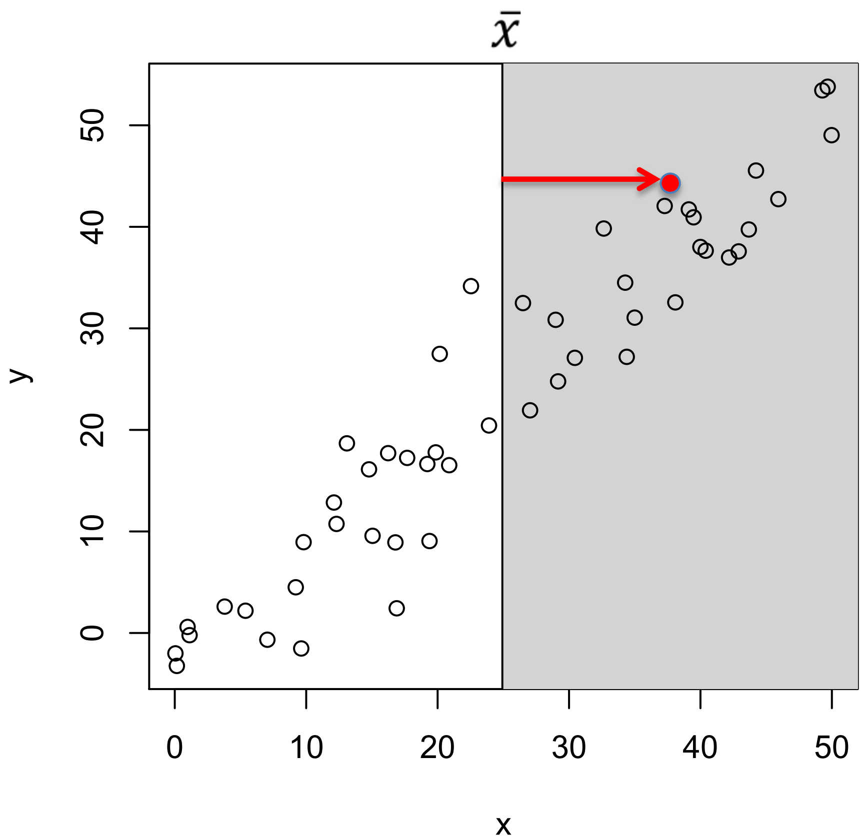
Correlation:



Correlation:

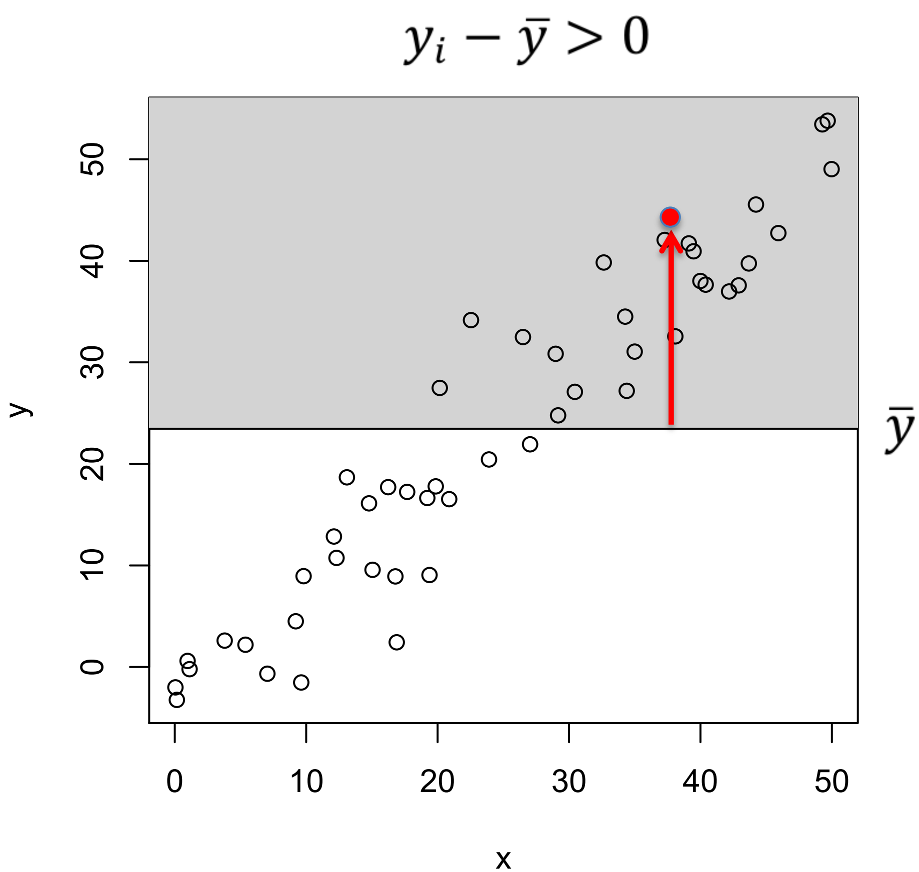


Correlation:

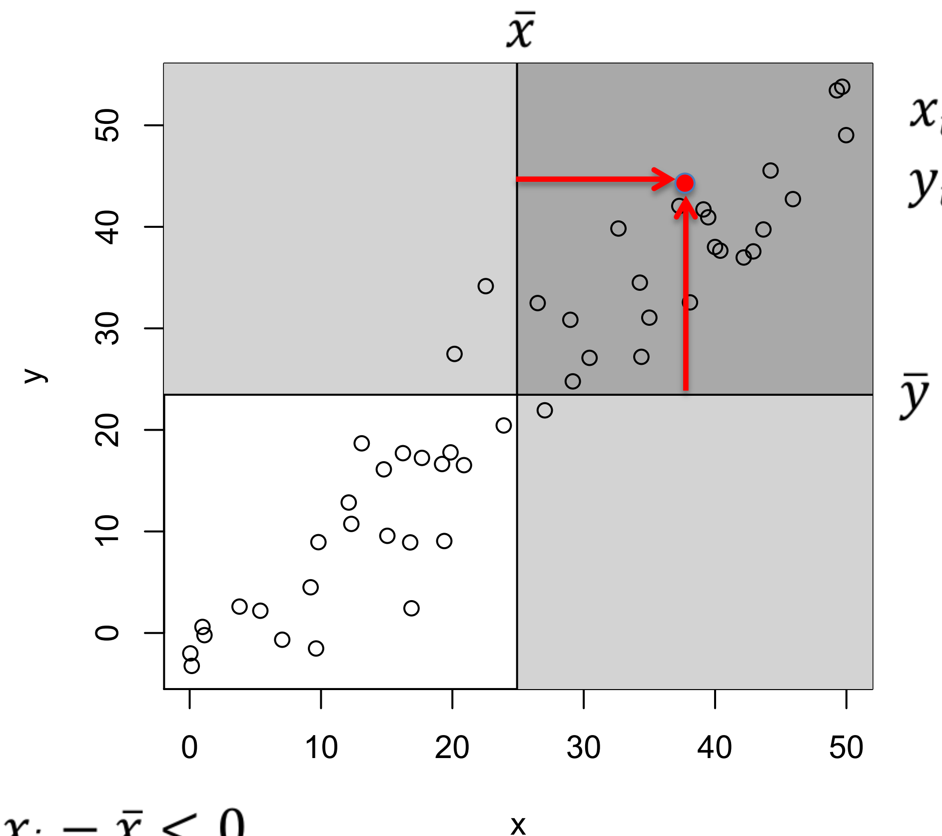


$$x_i - \bar{x} > 0$$

Correlation:



Correlation:



$x_i - \bar{x} > 0$

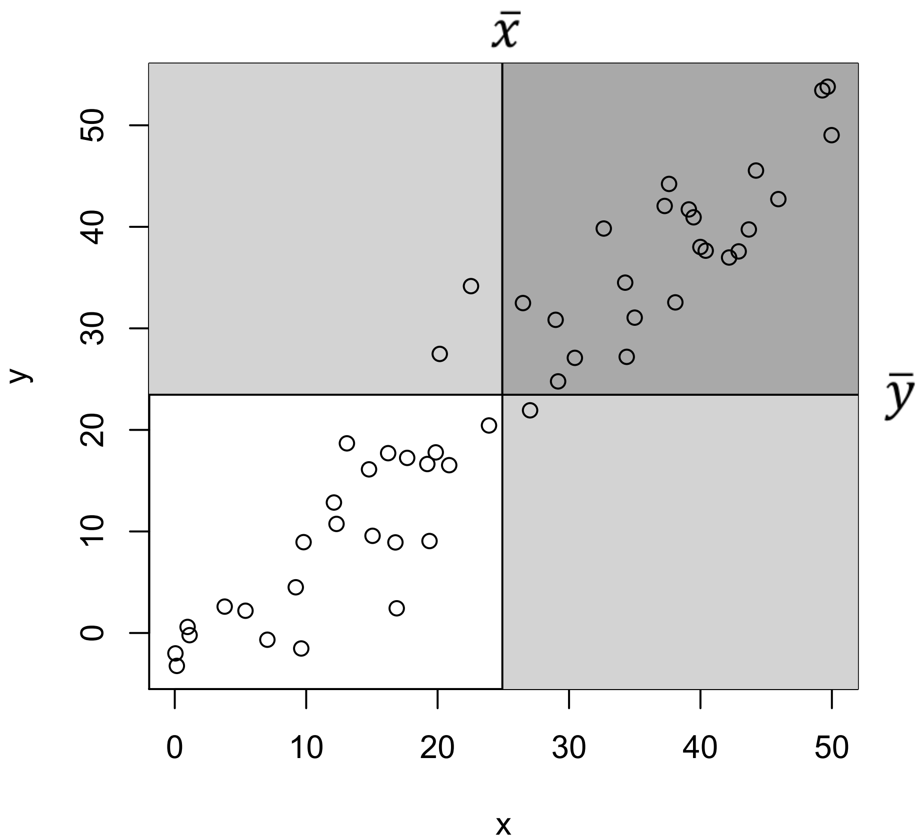
$y_i - \bar{y} > 0$

$x_i - \bar{x} < 0$

$y_i - \bar{y} < 0$

Correlation:

$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$

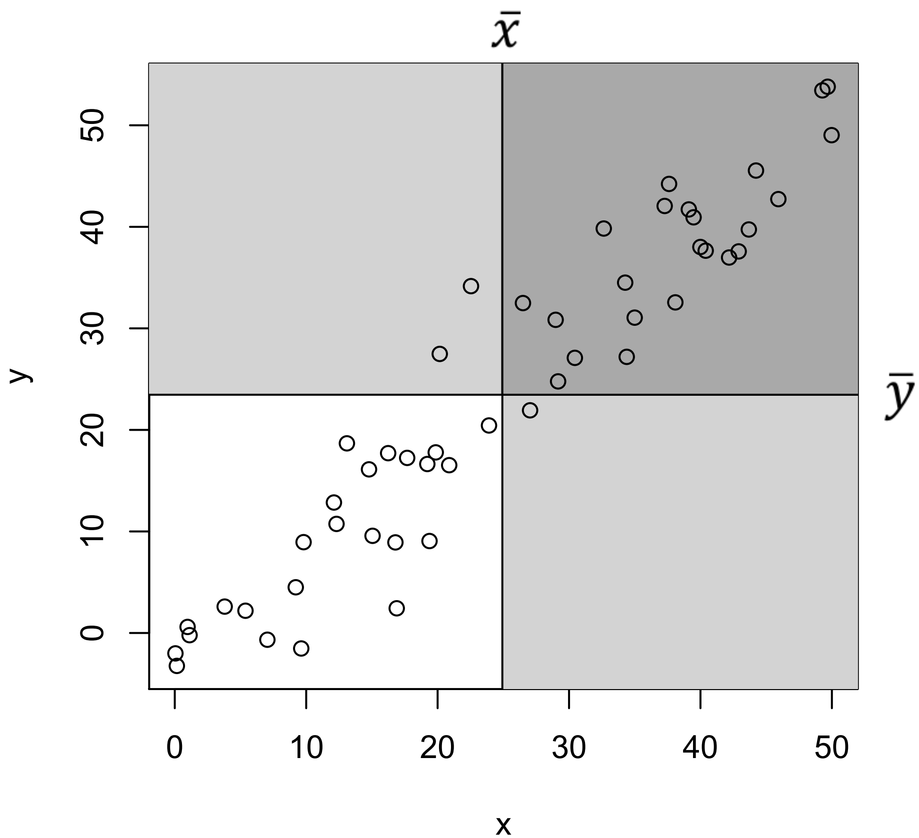


$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$

Correlation:

$(x_i - \bar{x})(y_i - \bar{y}) < 0$

$(x_i - \bar{x})(y_i - \bar{y}) > 0$



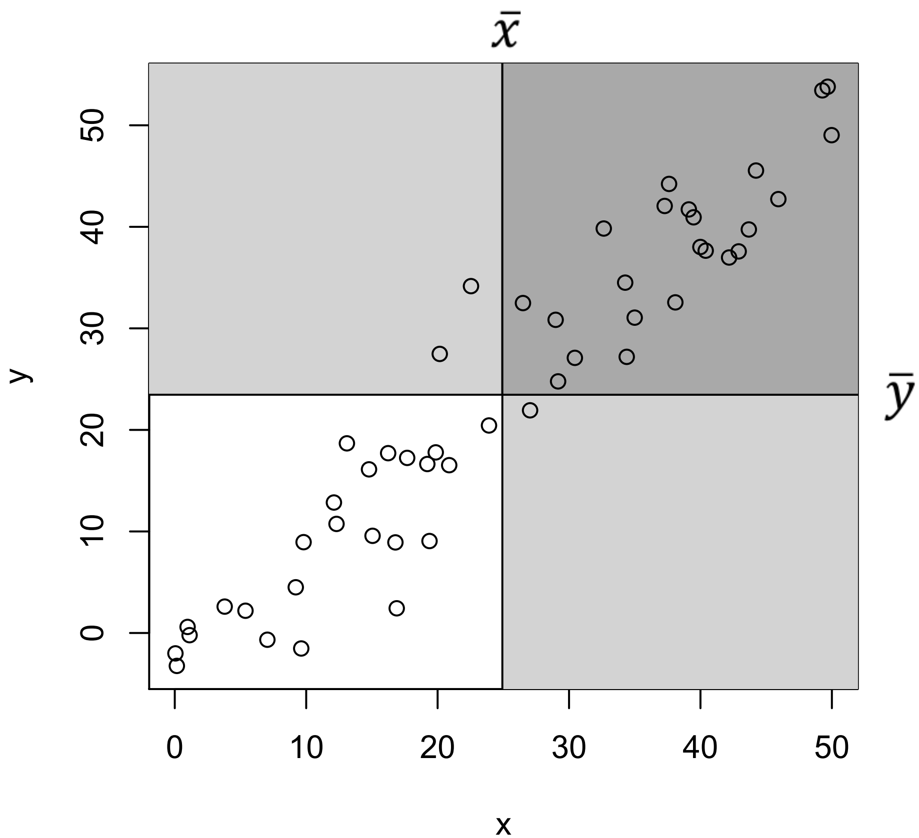
$(x_i - \bar{x})(y_i - \bar{y}) > 0$

$(x_i - \bar{x})(y_i - \bar{y}) < 0$

Correlation:

$(x_i - \bar{x})(y_i - \bar{y}) < 0$

$(x_i - \bar{x})(y_i - \bar{y}) > 0$



Positively correlated:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) \gg 0$$

Negatively correlated:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) \ll 0$$

Uncorrelated:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) \approx 0$$

$(x_i - \bar{x})(y_i - \bar{y}) > 0$

$(x_i - \bar{x})(y_i - \bar{y}) < 0$

Correlation:

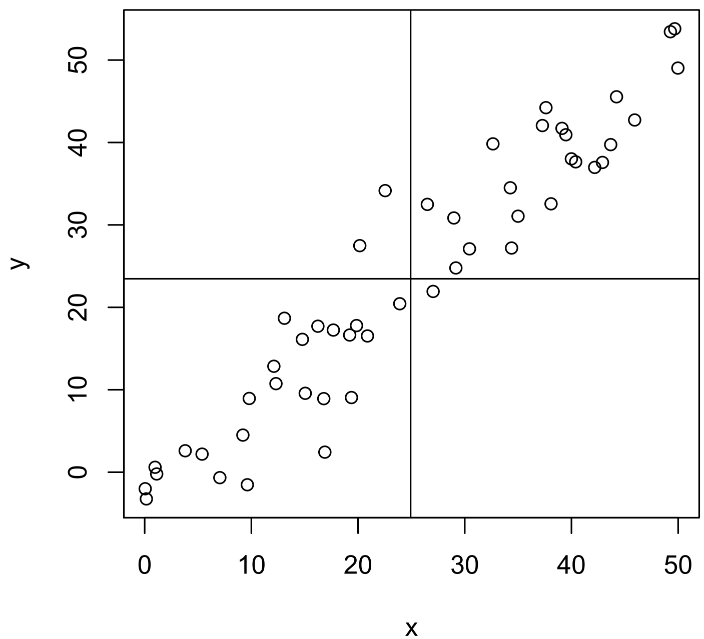
Pearson's product-moment correlation coefficient:

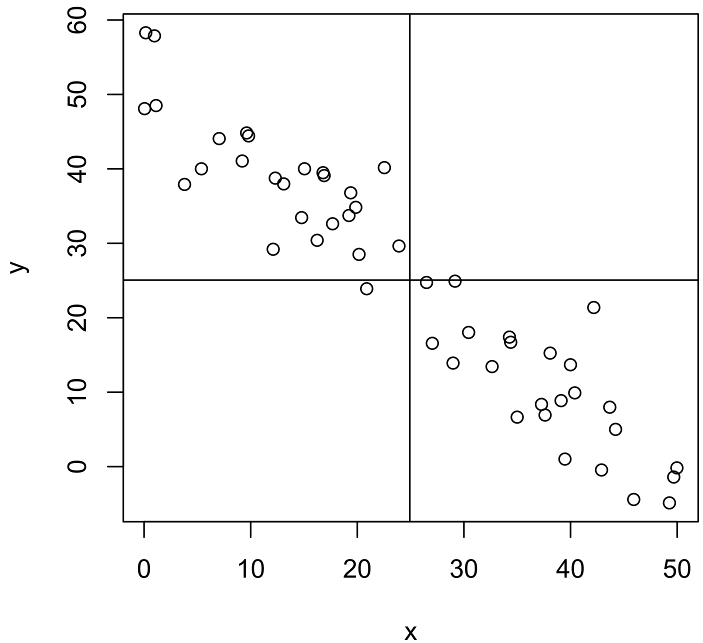
$$r_{X,Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Coefficient of Variation (R^2 value):

$$R^2_{X,Y} = {r_{X,Y}}^2$$

Correlation:





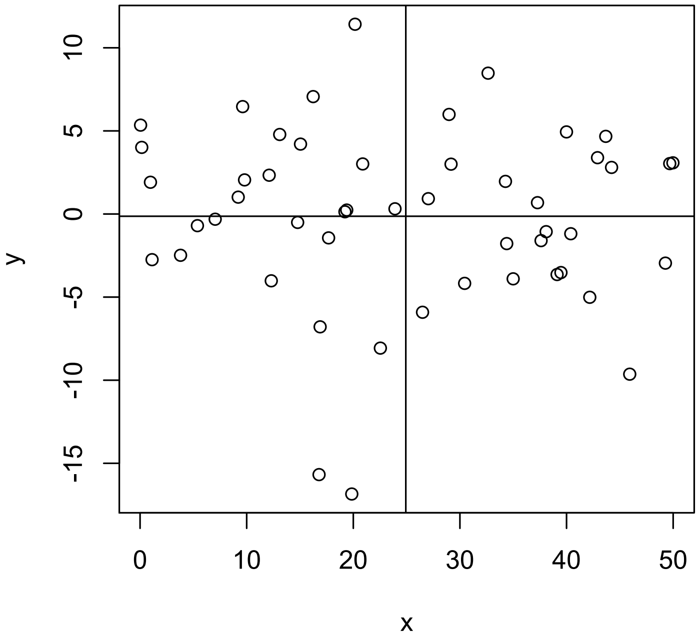
r = 0.931

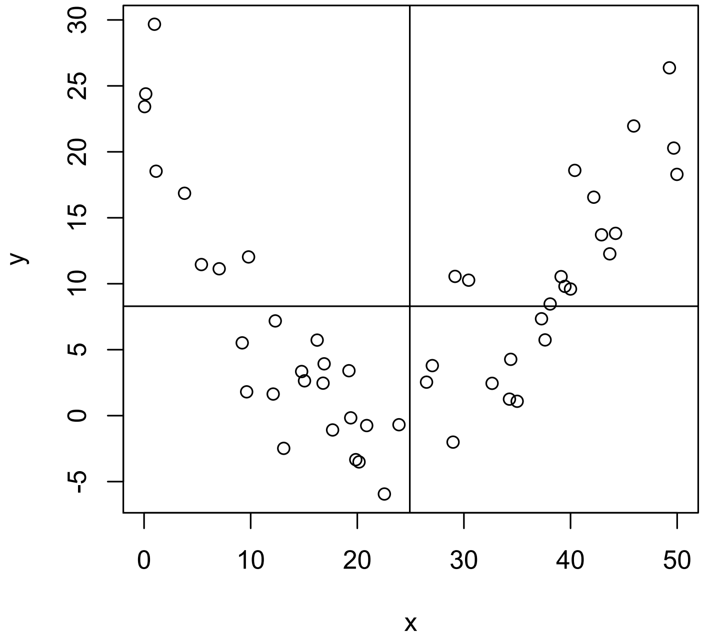
R^2 = 0.866

r = -0.949

R^2 = 0.901

Correlation:





r = -0.060

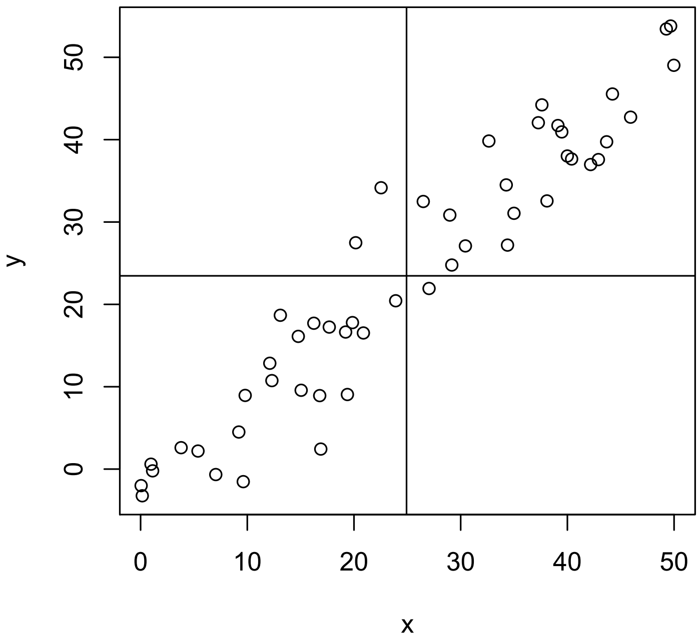
$R^2 = 0.004$

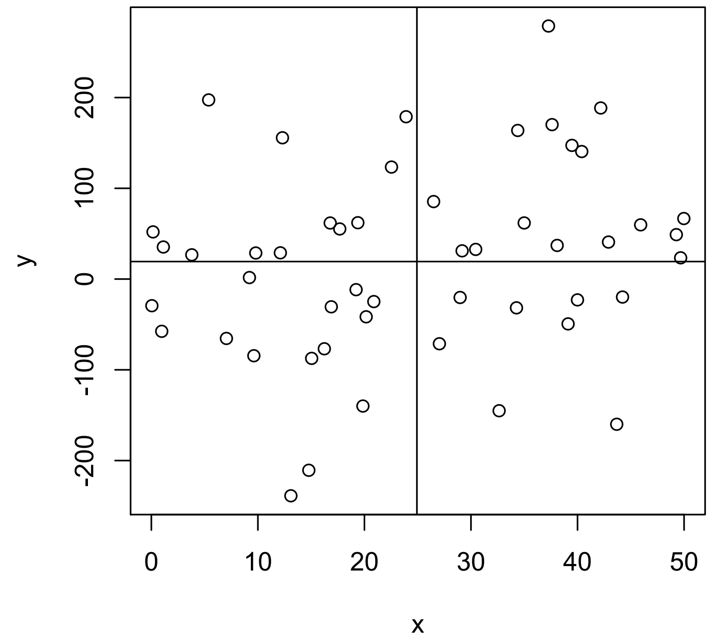
r = 0.106

$R^2 = 0.011$

Correlation:

Can I say whether my data are correlated?
Is an observed correlation significant?





data: x and y
t = 17.613, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8802556 0.9602168
sample estimates:
cor
0.9305923

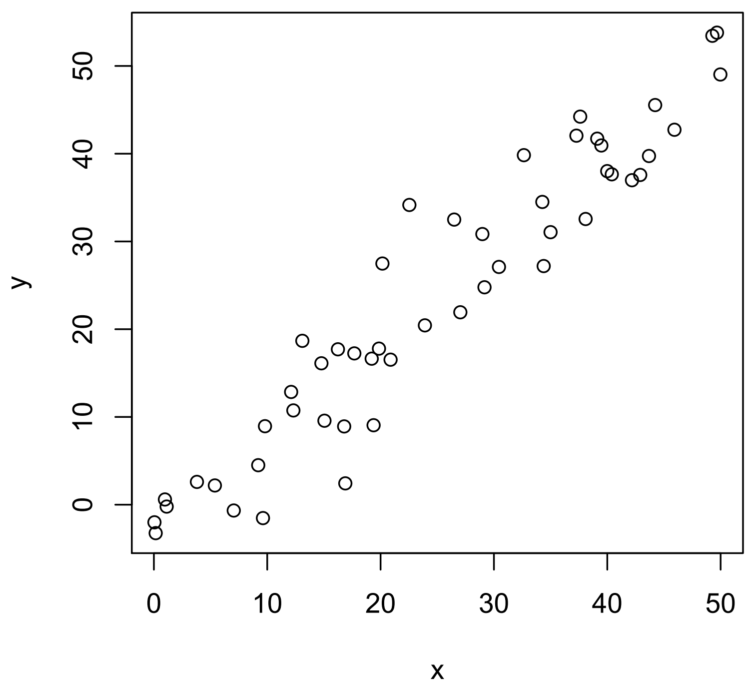
data: x and y
t = 1.5609, df = 48, p-value = 0.1251
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.06238066 0.46941403
sample estimates:
cor
0.2197833

Simple Regression:

Aims:

- To investigate linear correlation between two variables in more detail
- Be able to predict a response given knowledge of the independent variable

Response variable
Dependent variable



Predictor variable
Independent variable

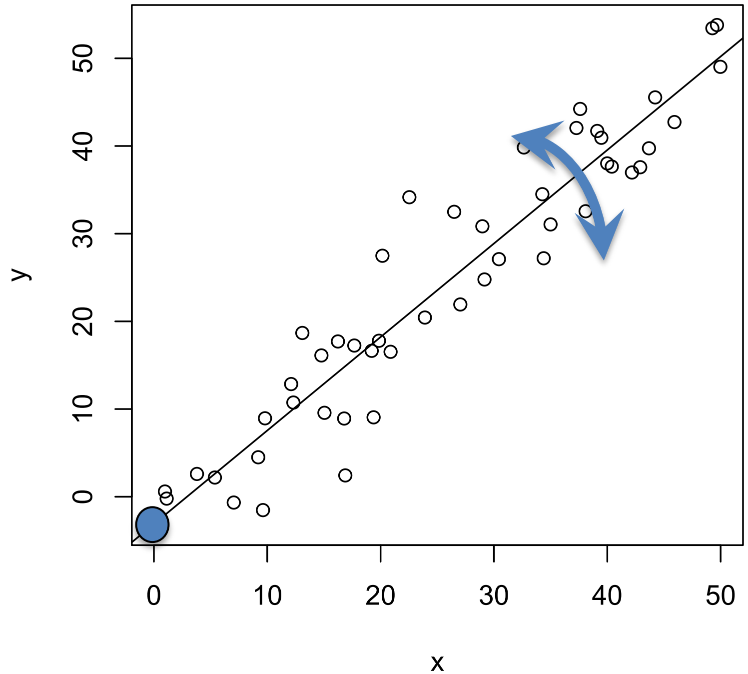
Simple Regression:

Aims:

- To investigate linear correlation between two variables in more detail
- Be able to predict a response given knowledge of the independent variable

$$y = \alpha + \beta x$$

Response variable
Dependent variable



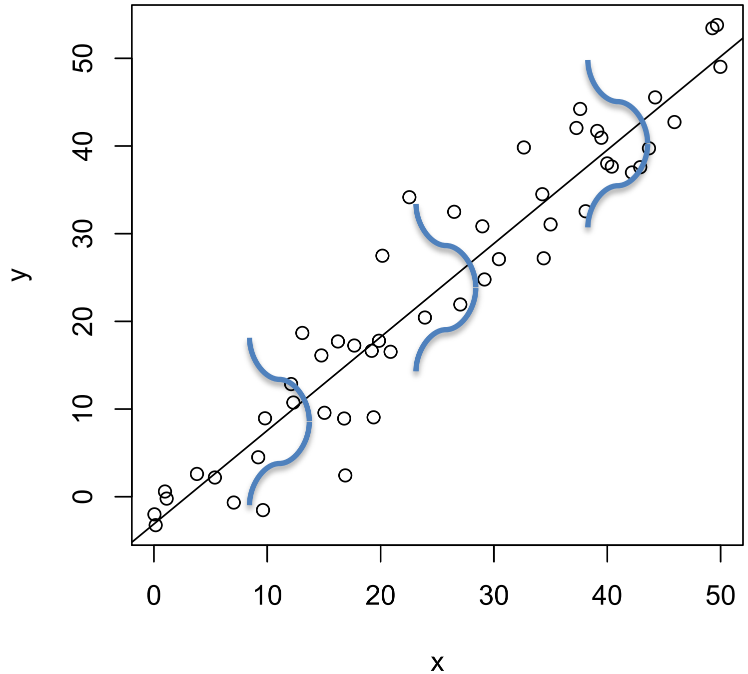
Predictor variable
Independent variable

Simple Regression:

Aims:

- To investigate linear correlation between two variables in more detail
- Be able to predict a response given knowledge of the independent variable

Response variable
Dependent variable



$$y = \alpha + \beta x + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

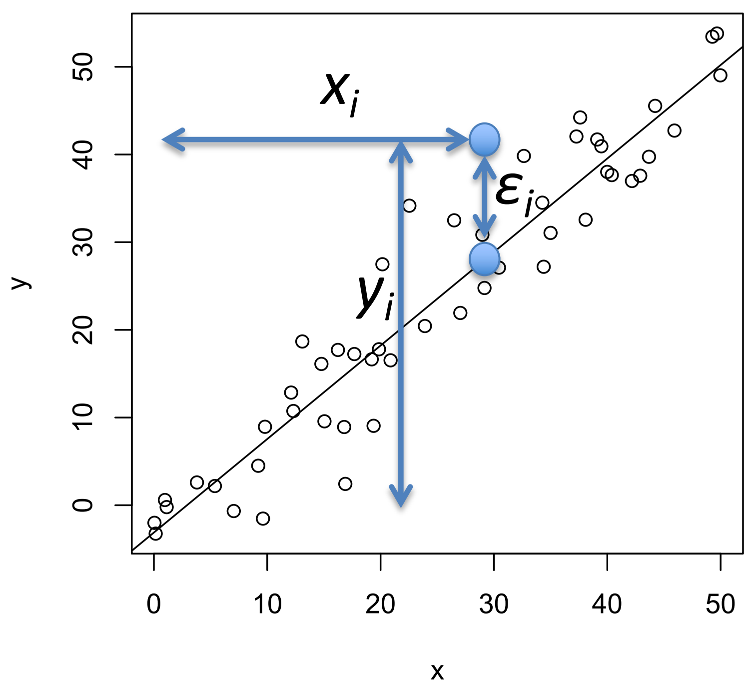
Predictor variable
Independent variable

Simple Regression:

Aims:

- To investigate linear correlation between two variables in more detail
- Be able to predict a response given knowledge of the independent variable

Response variable
Dependent variable



Predictor variable
Independent variable

$$y = \alpha + \beta x + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

For the i^{th} observation:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

ε_i = errors, residuals

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

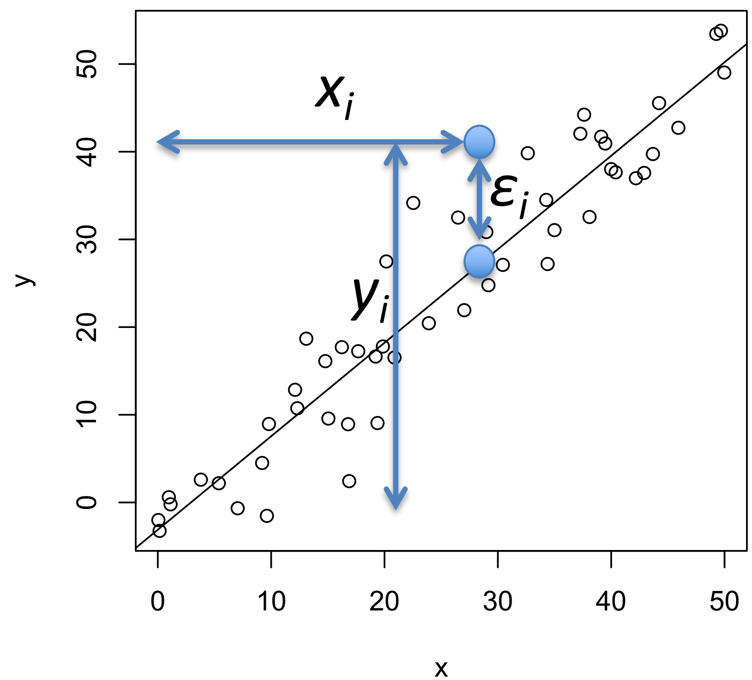
$$\varepsilon_i \sim N(0, \sigma^2)$$

Simple Regression:

So how do we fit the regression line?

Suppose we know parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

Observations: $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$



Simple Regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

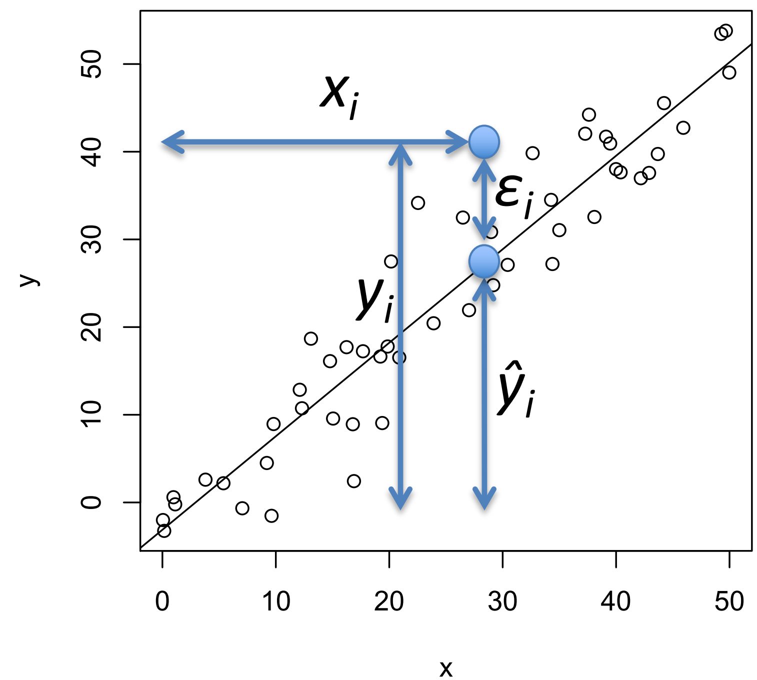
$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

So how do we fit the regression line?

Suppose we know parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

Observations: $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$

Fitted values:
$$\begin{aligned}\hat{\boldsymbol{y}} &= E(\boldsymbol{y} \mid \boldsymbol{x};\ \hat{\alpha}, \hat{\beta}) \\ &= E(\alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon} | \hat{\alpha}, \hat{\beta}) \\ &= E(\hat{\alpha} + \hat{\beta} \boldsymbol{x} + \boldsymbol{\varepsilon}) \\ &= \hat{\alpha} + \hat{\beta} \boldsymbol{x}\end{aligned}$$



$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

Simple Regression:

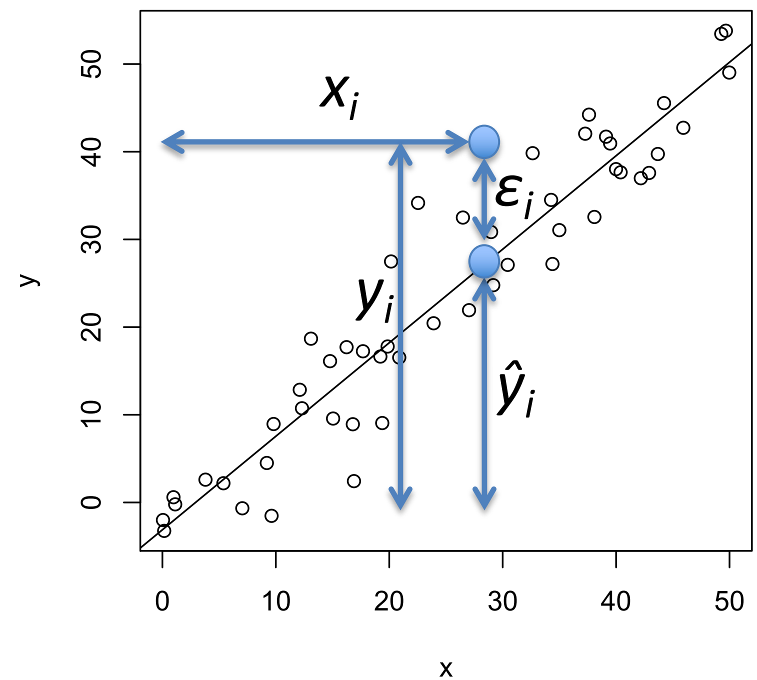
$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta} \boldsymbol{x}$$

So how do we fit the regression line?

Suppose we know parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

Residuals: $\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$



Simple Regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\widehat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta} \boldsymbol{x}$$

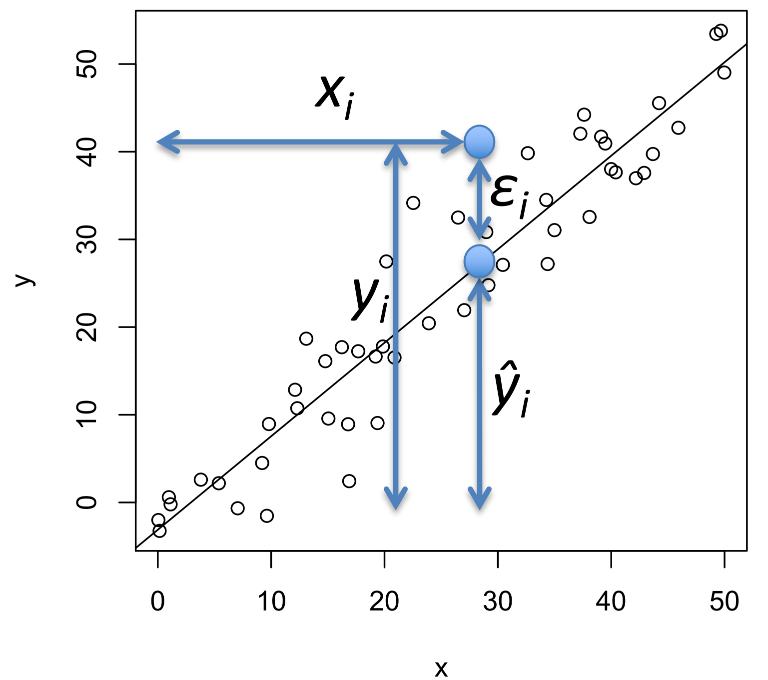
$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

So how do we fit the regression line?

Suppose we know parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

Residuals: $\boldsymbol{\varepsilon} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$

$$\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{\varepsilon}$$



Simple Regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\widehat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta} \boldsymbol{x}$$

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

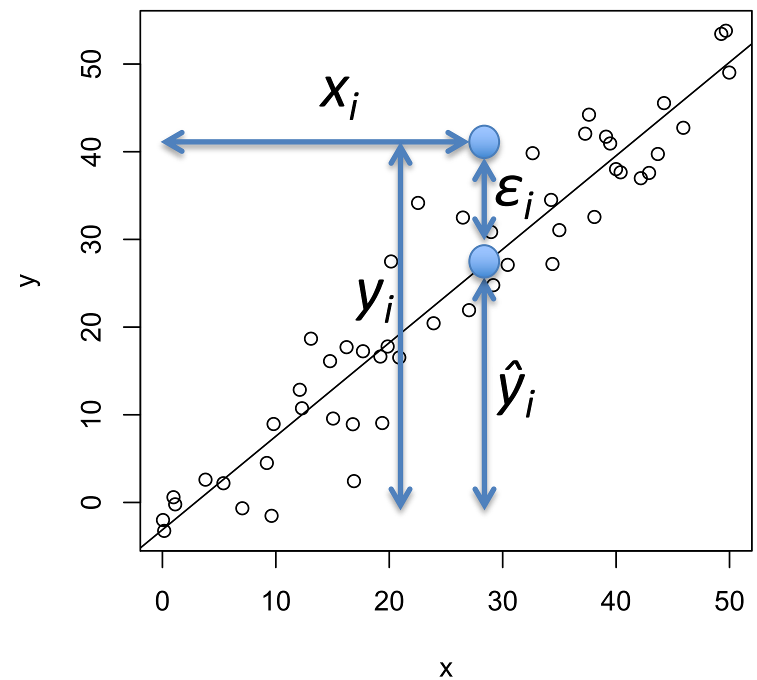
So how do we fit the regression line?

Suppose we know parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

Residuals: $$\boldsymbol{\varepsilon} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

$$\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} \sim N(\widehat{\boldsymbol{y}}, \sigma^2)$$



Simple Regression:

So how do we fit the regression line?

Suppose we know parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\widehat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta} \boldsymbol{x}$$

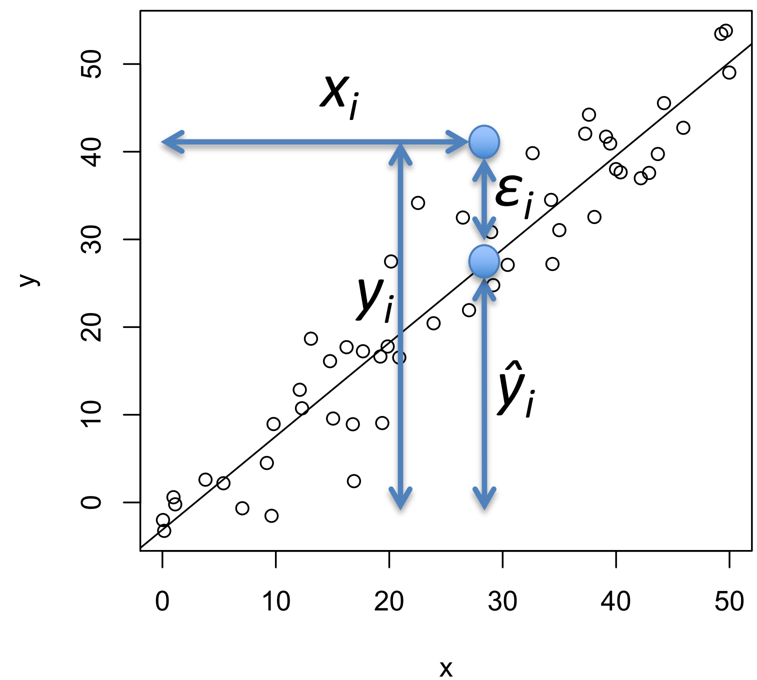
$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

Residuals: $\boldsymbol{\varepsilon} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$

$$\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} \sim N(\widehat{\boldsymbol{y}}, \sigma^2)$$

$$f(\boldsymbol{y}|\boldsymbol{x}; \hat{\alpha}, \hat{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\boldsymbol{y}-\widehat{\boldsymbol{y}})^2}{2\sigma^2}}$$



Simple Regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\widehat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta}\boldsymbol{x}$$

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

So how do we fit the regression line?

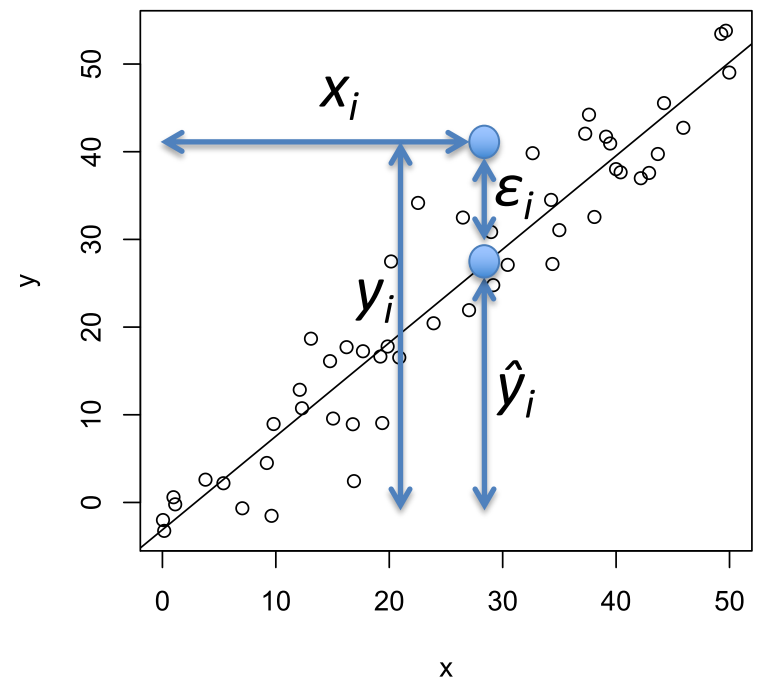
Obtain estimates $\hat{\alpha}$ and $\hat{\beta}$

Maximise likelihood of parameters given the data

$$f(\boldsymbol{y}|\boldsymbol{x}; \hat{\alpha}, \hat{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\boldsymbol{y}-\widehat{\boldsymbol{y}})^2}{2\sigma^2}}$$

$$\mathcal{L}(\hat{\alpha}, \hat{\beta}|\boldsymbol{y}, \boldsymbol{x}) = \prod_i f(y_i|x_i; \hat{\alpha}, \hat{\beta})$$

$$= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\hat{y}_i)^2}{2\sigma^2}}$$



Simple Regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta} \boldsymbol{x}$$

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

So how do we fit the regression line?

Obtain estimates $\hat{\alpha}$ and $\hat{\beta}$

Maximise likelihood of parameters given the data

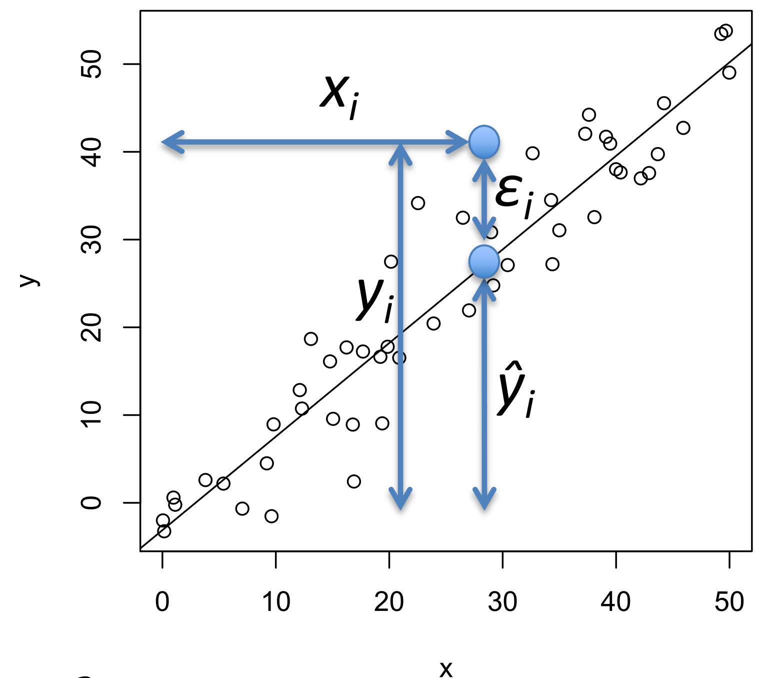
$$f(\boldsymbol{y}|\boldsymbol{x}; \hat{\alpha}, \hat{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\boldsymbol{y}-\hat{\boldsymbol{y}})^2}{2\sigma^2}}$$

$$\mathcal{L}(\hat{\alpha}, \hat{\beta}|\boldsymbol{y}, \boldsymbol{x}) = \prod_i f(y_i|x_i; \hat{\alpha}, \hat{\beta})$$

$$= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\hat{y}_i)^2}{2\sigma^2}}$$

$$\ln \mathcal{L}(\hat{\alpha}, \hat{\beta}|\boldsymbol{y}, \boldsymbol{x}) = \sum_i \left(\frac{-1}{2} \ln(2\pi\sigma^2) - \frac{(y_i-\hat{y}_i)^2}{2\sigma^2} \right)$$

$$= \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \hat{y}_i)^2$$



Simple Regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta} \boldsymbol{x}$$

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

So how do we fit the regression line?

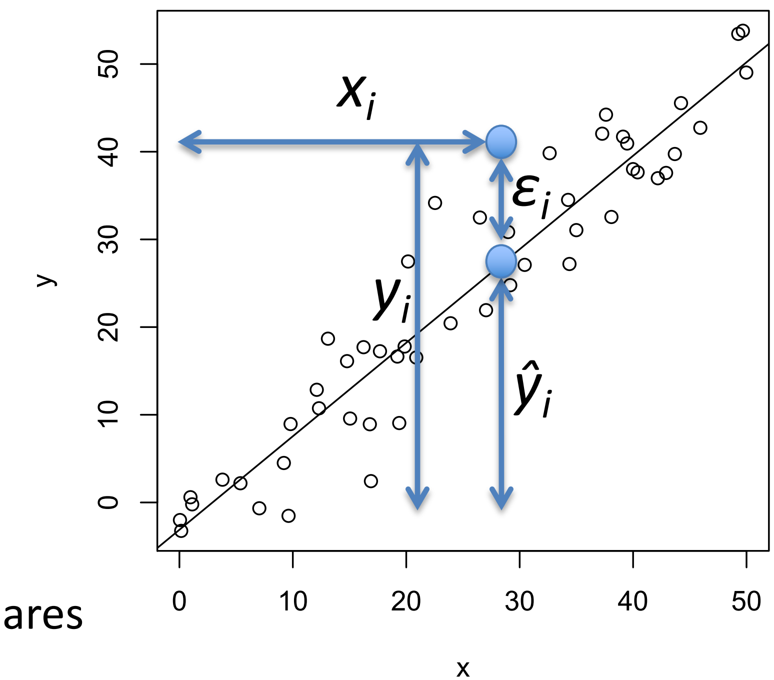
Obtain estimates $\hat{\alpha}$ and $\hat{\beta}$

Maximise likelihood of parameters given the data

$$\mathcal{L}(\hat{\alpha}, \hat{\beta} \,|\, \boldsymbol{y}, \boldsymbol{x}) \rightarrow max$$

$$\sum_i (y_i - \hat{y}_i)^2 \rightarrow min$$

$$\sum_i {\varepsilon_i}^2 \rightarrow min$$



Optimal parameters : minimise residual sum of squares

Maximum Likelihood and Least Squares estimates are equivalent (for Gaussian error model)

Simple Regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta} \boldsymbol{x}$$

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

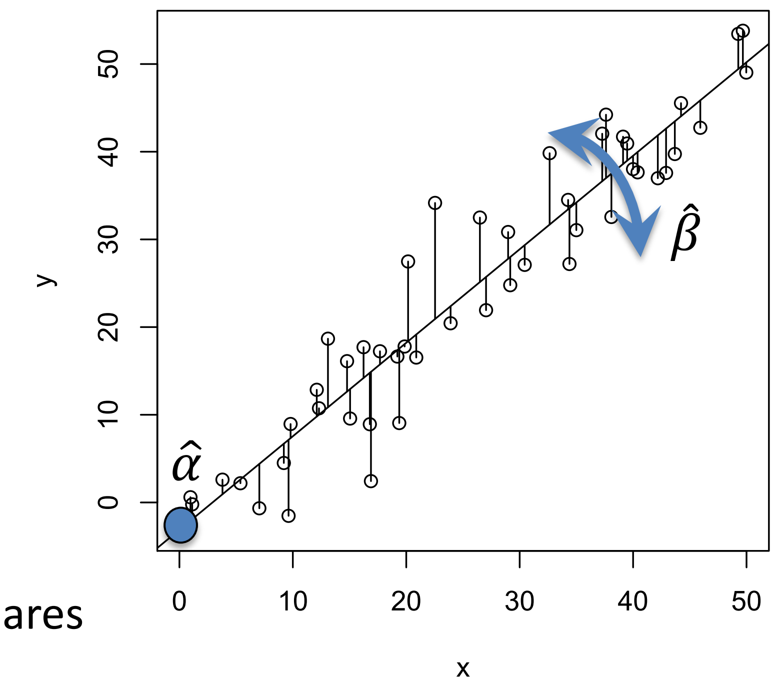
So how do we fit the regression line?

Obtain estimates $\hat{\alpha}$ and $\hat{\beta}$
Maximise likelihood of parameters given the data

$$\mathcal{L}(\hat{\alpha}, \hat{\beta} \,|\, \boldsymbol{y}, \boldsymbol{x}) \to max$$

$$\sum_i (y_i - \hat{y}_i)^2 \to min$$

$$\sum_i {\varepsilon_i}^2 \to min$$



Optimal parameters : minimise residual sum of squares

Maximum Likelihood and Least Squares estimates are equivalent (for Gaussian error model)

Simple Regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta} \boldsymbol{x}$$

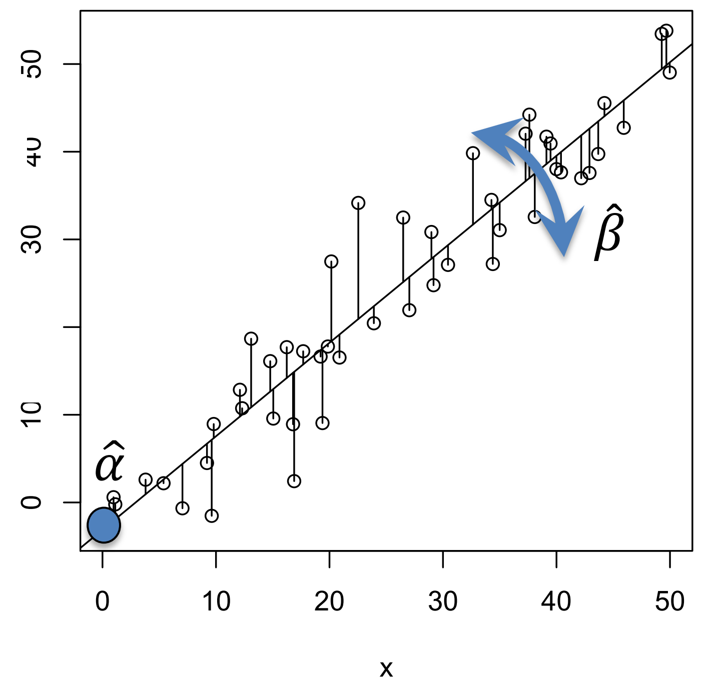
$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

So how do we fit the regression line?

Obtain estimates $\hat{\alpha}$ and $\hat{\beta}$
Maximise likelihood of parameters given the data
Minimise sum of squared residuals

$$\sum_i \varepsilon_i^2 \rightarrow min$$

$$\sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \xrightarrow[\hat{\alpha}, \hat{\beta}]{} min$$



Final answer:

$$\hat{\beta} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\text{cov}(\boldsymbol{x}, \boldsymbol{y})}{\text{var}(\boldsymbol{x})}$$

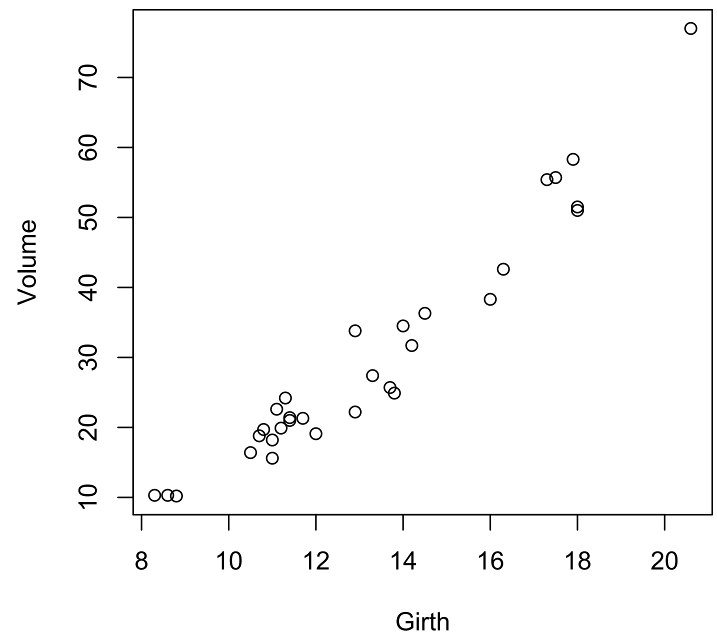
$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Simple Regression:

Example: *Predicting timber volume of felled black cherry trees*

```
> cor(trees$Volume,trees$Girth)
[1] 0.9671194
```

```
> m1 = lm(Volume~Girth,data=trees)
> summary(m1)

Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min     1Q Median     3Q    Max 
-8.065 -3.107  0.152  3.495  9.587 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,     Adjusted R-squared:  0.9331 
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```



Response: y = Volume

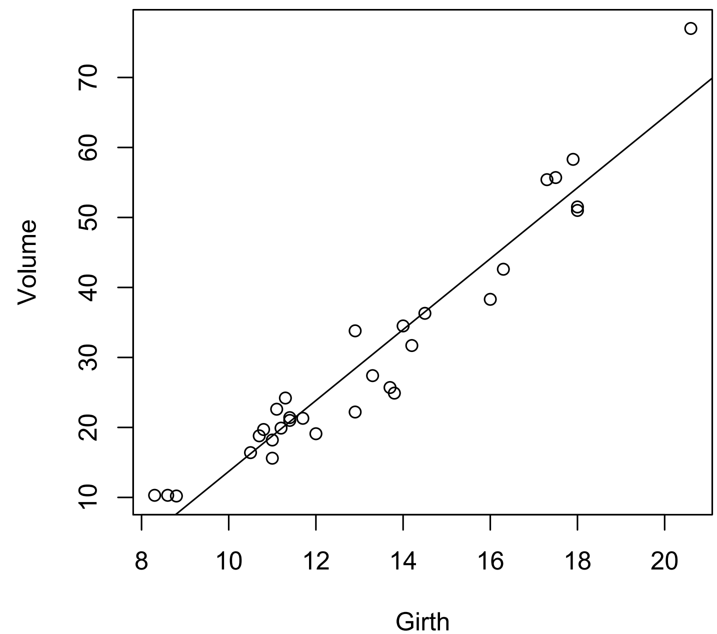
Predictor: x = Girth

Simple Regression:

Example: *Predicting timber volume of felled black cherry trees*

```
> cor(trees$Volume,trees$Girth)
[1] 0.9671194
```

```
> m1 = lm(Volume~Girth,data=trees)
> summary(m1)

Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min     1Q Median     3Q    Max 
-8.065 -3.107  0.152  3.495  9.587 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,	Adjusted R-squared:  0.9331 
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```



Response: y = Volume
Predictor: x = Girth

$$y = -36.9 + 5.07x$$

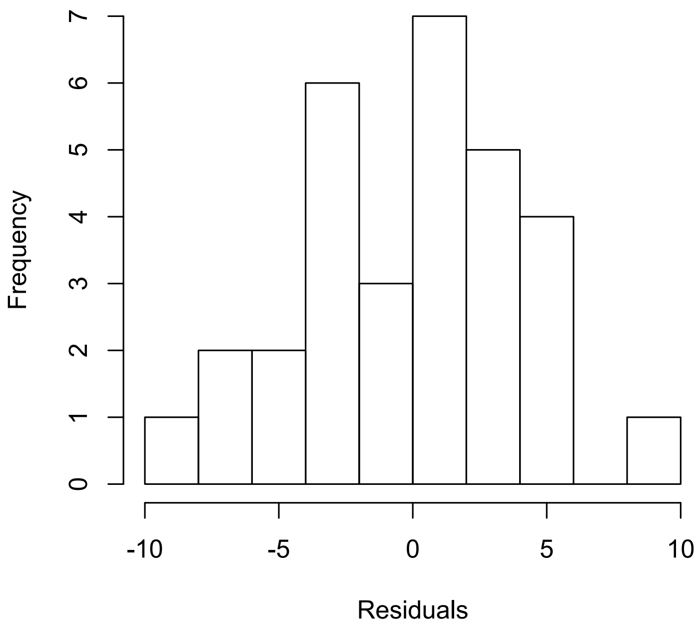
Simple Regression:

Example: *Predicting timber volume of felled black cherry trees*

```
> cor(trees$Volume,trees$Girth)
[1] 0.9671194
```

```
> m1 = lm(Volume~Girth,data=trees)
> summary(m1)

Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min     1Q Median     3Q    Max 
-8.065 -3.107  0.152  3.495  9.587 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,     Adjusted R-squared:  0.9331 
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

$$\sigma = 4.252$$
$$\sigma^2 = 18.1$$



Response: y = Volume
Predictor: x = Girth

$$y = -36.9 + 5.07x + \varepsilon$$
$$\varepsilon \sim N(0,18.1)$$

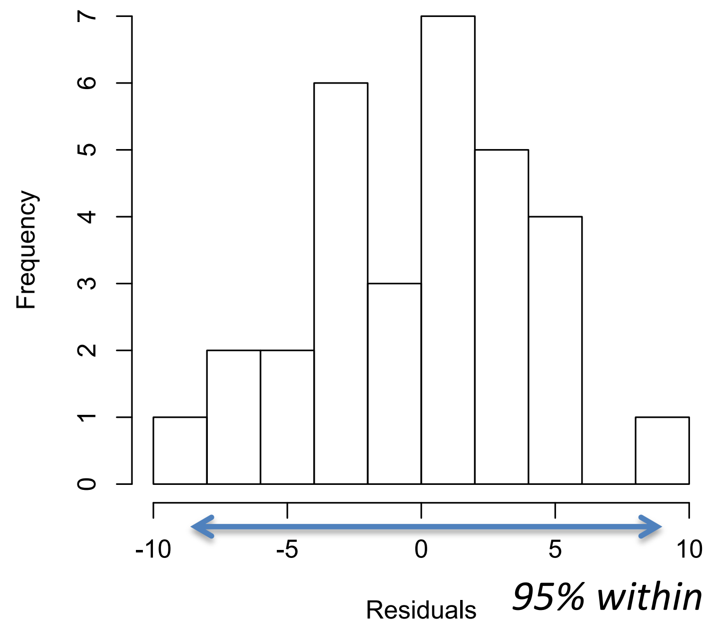
Simple Regression:

Example: *Predicting timber volume of felled black cherry trees*

```
> cor(trees$Volume,trees$Girth)
[1] 0.9671194
```

```
> m1 = lm(Volume~Girth,data=trees)
> summary(m1)

Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min     1Q Median     3Q    Max 
-8.065 -3.107  0.152  3.495  9.587 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,	Adjusted R-squared:  0.9331 
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

$$\sigma = 4.252$$
$$\sigma^2 = 18.1$$



95% within ±8.5

Response: y = Volume

Predictor: x = Girth

$$y = -36.9 + 5.07x + \varepsilon$$

$$\varepsilon \sim N(0,18.1)$$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x}^2 + \boldsymbol{\varepsilon}$$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x}^2 + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \alpha + \boldsymbol{x}^{\beta} + \boldsymbol{\varepsilon}$$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x}^2 + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \alpha + \boldsymbol{x}^{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \alpha + \beta \log(\boldsymbol{x}) + \boldsymbol{\varepsilon}$$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x}^2 + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \alpha + \boldsymbol{x}^\beta + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \alpha + \beta \log(\boldsymbol{x}) + \boldsymbol{\varepsilon}$$

$$\log(\boldsymbol{y}) = \alpha + \beta \sqrt{\boldsymbol{x}} + \boldsymbol{\varepsilon}$$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$

2. Gaussian error model. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon}$$

2. Gaussian error model. $$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

3. Additive error model. $$\boldsymbol{y} = \alpha + \beta\boldsymbol{x}\boldsymbol{\varepsilon}$$ $$\boldsymbol{y} = \alpha + \beta\boldsymbol{x}^{\boldsymbol{\varepsilon}}$$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

2. Gaussian error model. $$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

3. Additive error model. $$\cancel{\boldsymbol{y} = \alpha + \beta \boldsymbol{x} \boldsymbol{\varepsilon}}$$ $$\cancel{\boldsymbol{y} = \alpha + \beta \boldsymbol{x}^{\boldsymbol{\varepsilon}}}$$

Linear Regression:

Assumptions:

1. Model is linear in parameters. $\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$
2. Gaussian error model. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$
3. Additive error model. $\cancel{\boldsymbol{y} = \alpha + \beta \boldsymbol{x} \boldsymbol{\varepsilon}}$ $\cancel{\boldsymbol{y} = \alpha + \beta \boldsymbol{x}^{\boldsymbol{\varepsilon}}}$
4. Independence of errors. $\mathrm{Cov}\left(\varepsilon_i, \varepsilon_j\right) = 0$

   No autocorrelation – when one observation depends on another

Linear Regression:

Assumptions:

1. Model is linear in parameters. $\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon}$

2. Gaussian error model. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

3. Additive error model.

$$\cancel{\boldsymbol{y} = \alpha + \beta\boldsymbol{x}\boldsymbol{\varepsilon}}$$
$$\cancel{\boldsymbol{y} = \alpha + \beta\boldsymbol{x}^{\varepsilon}}$$

4. Independence of errors. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$

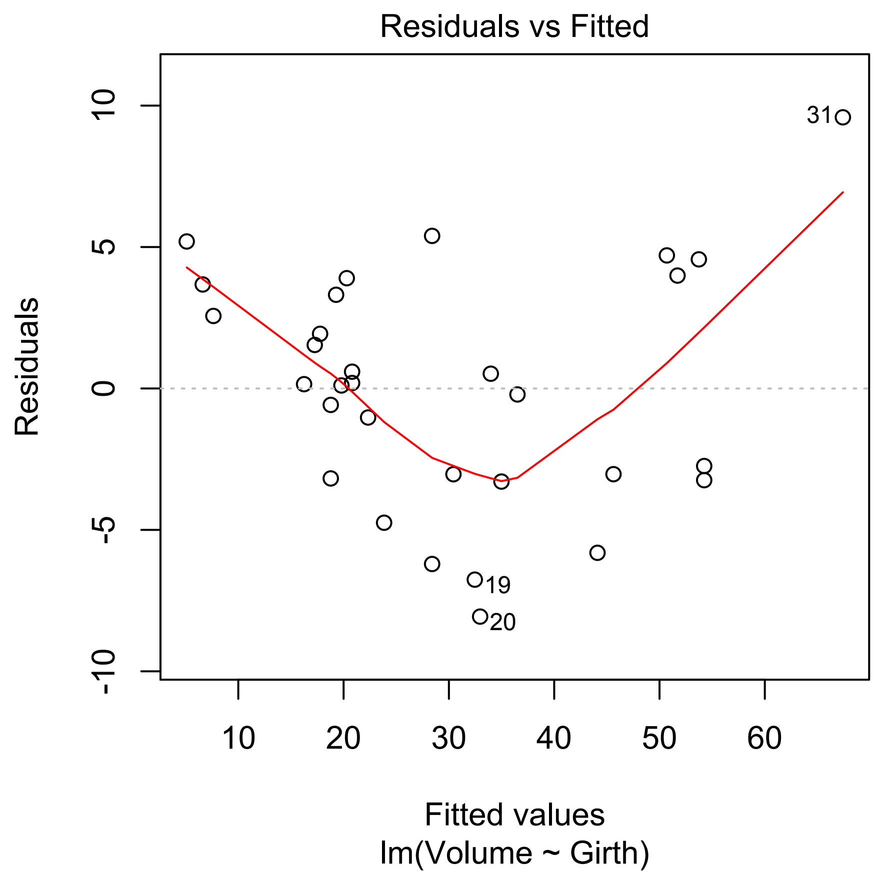
   No autocorrelation – when one observation depends on another

5. Homoscedasticity. $\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{x}) = \sigma^2\mathbf{I}$

   Homogeneity / stability of variance of the residuals

Testing Assumptions: diagnostic plots

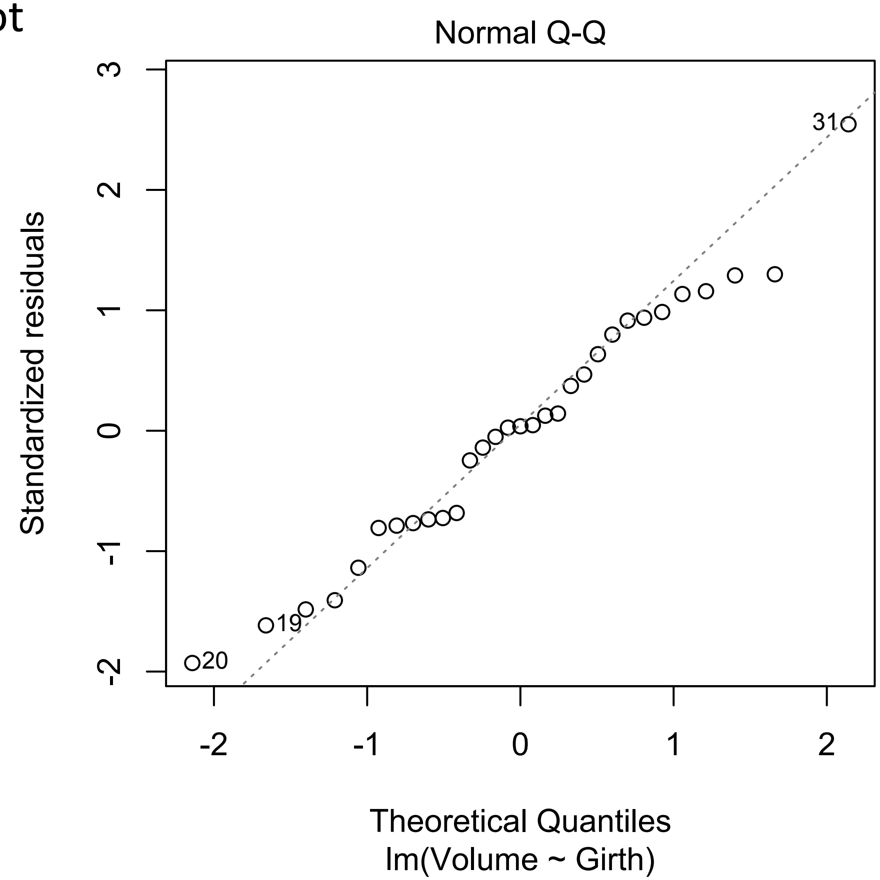
1. Residuals vs Fitted Values



- Should not be related
- No visible pattern
- Mean residual = zero
- Constant variance

Testing Assumptions: diagnostic plots

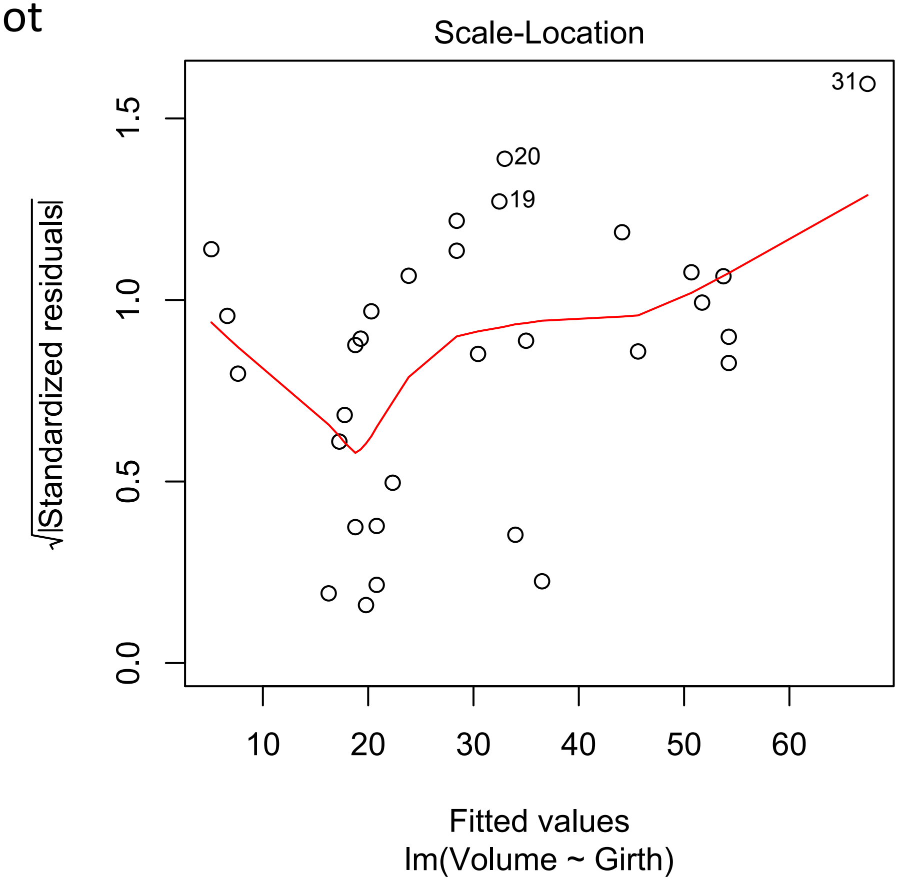
1. Residuals vs Fitted Values
2. Normal Quantile-Quantile Plot



- Visual test for Normality
- No strong trends/departures

Testing Assumptions: diagnostic plots

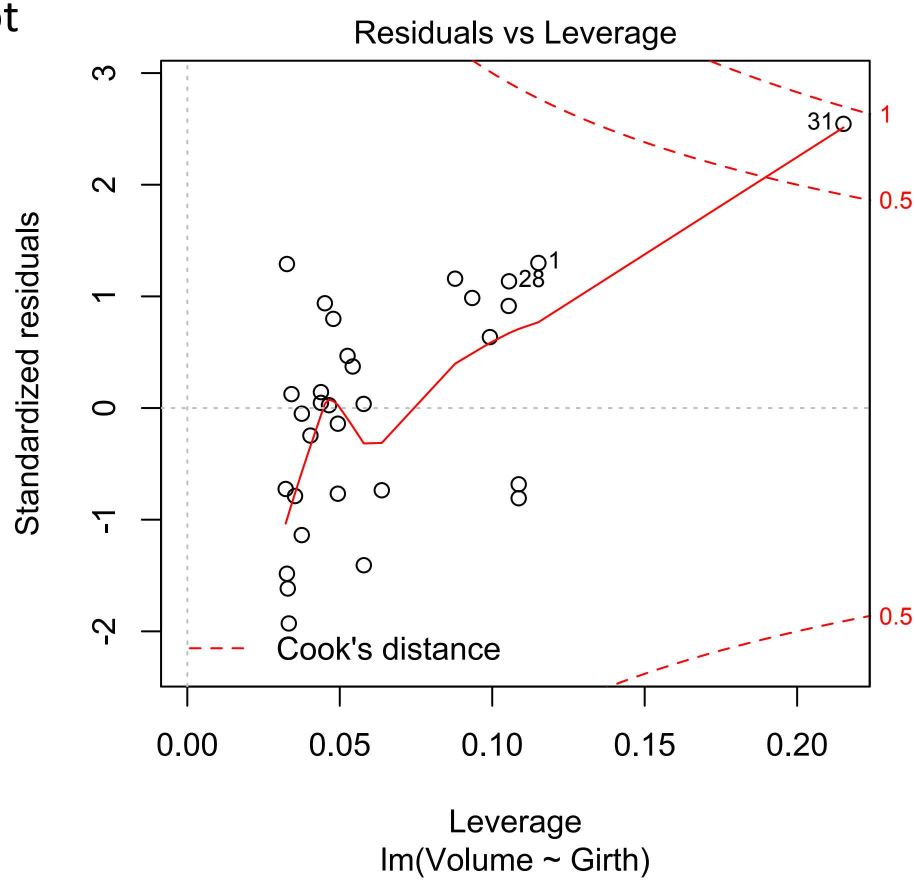
1. Residuals vs Fitted Values
2. Normal Quantile-Quantile Plot
3. Scale-Location Plot



- Test for homoscedasticity
- Should be constant, ≈1
- No trend

Testing Assumptions: diagnostic plots

1. Residuals vs Fitted Values
2. Normal Quantile-Quantile Plot
3. Scale-Location Plot
4. Index Plot of Cook's Distance



- Measures the influence of a particular observation
- Extreme x-vals : high leverage
- May inform outlier rejection

Modelling Non-Linear Relationships

Linear models can be used to describe non-linear relationships…

$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x}^2 + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \alpha + \beta\log(\boldsymbol{x}) + \boldsymbol{\varepsilon}$$

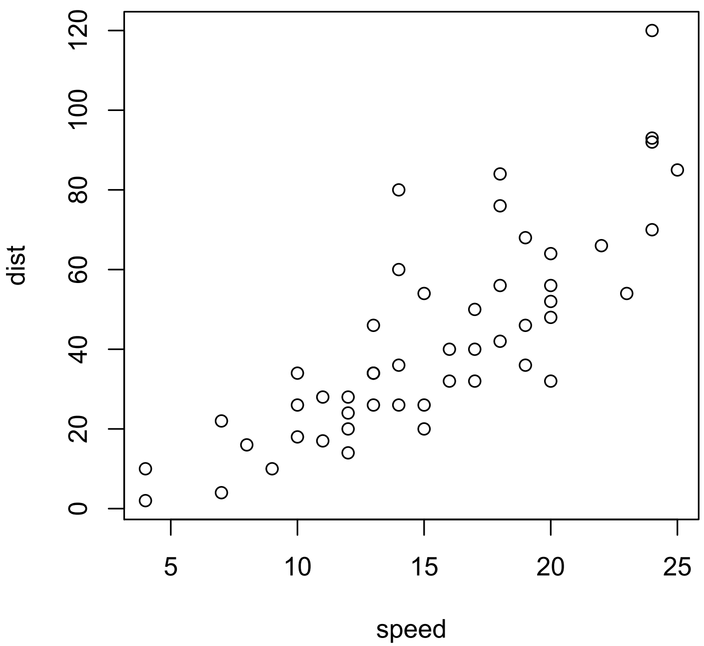
$$\log(\boldsymbol{y}) = \alpha + \beta\sqrt{\boldsymbol{x}} + \boldsymbol{\varepsilon}$$

Applying transformations to response and/or predictor variables can be useful to:

- Linearise the data, i.e. make the relationship between variables more linear.
- Stabilise the variance of the residuals, so that σ^2 doesn't depend on the independent variable.
- Normalise the distribution of the residuals

Modelling Non-Linear Relationships

Example: *Stopping distance of cars versus speed (mph)*



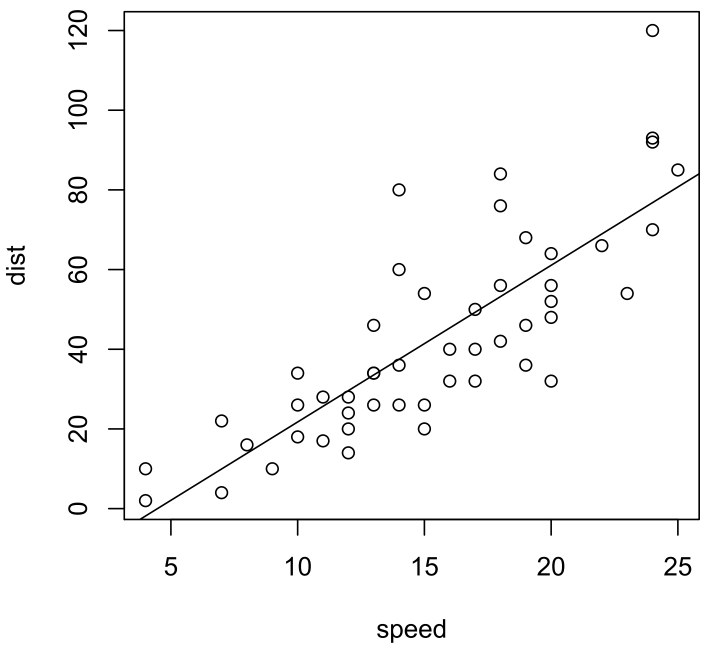
Response: y = distance
Predictor: x = speed

Modelling Non-Linear Relationships

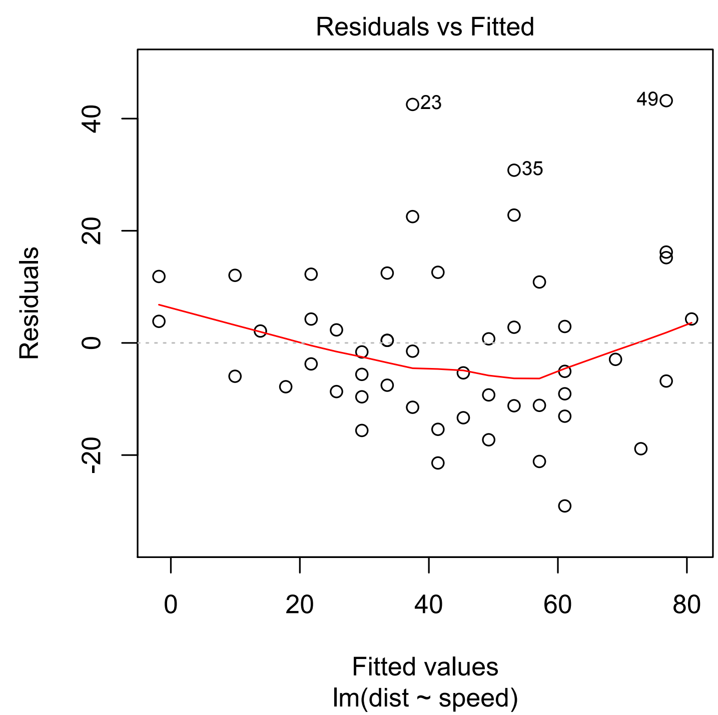
Example: *Stopping distance of cars versus speed (mph)*

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$R^2 = 0.651$

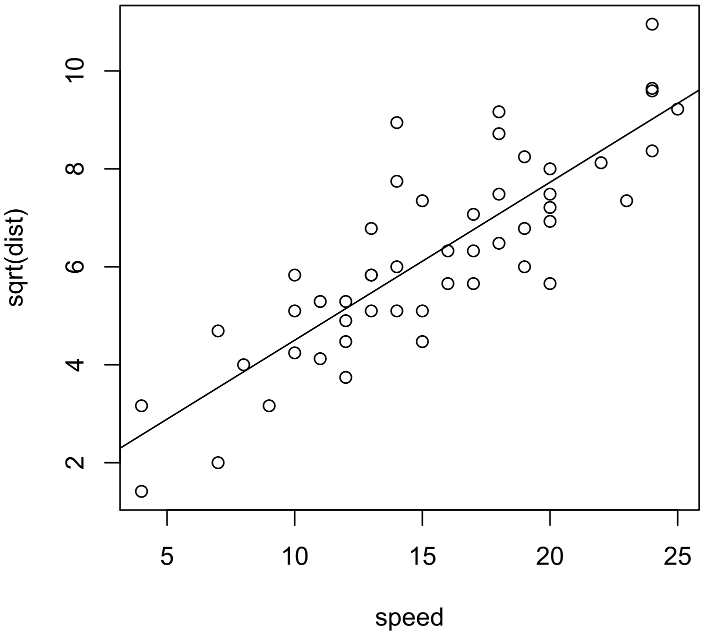


Response: y = distance
Predictor: x = speed



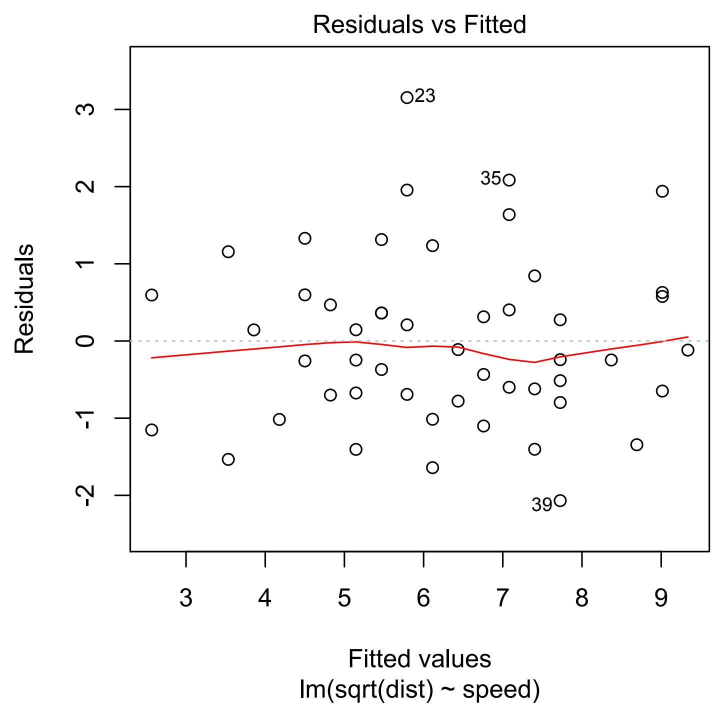
Modelling Non-Linear Relationships

Example: *Stopping distance of cars versus speed (mph)*



$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon} \qquad R^2 = 0.651$$

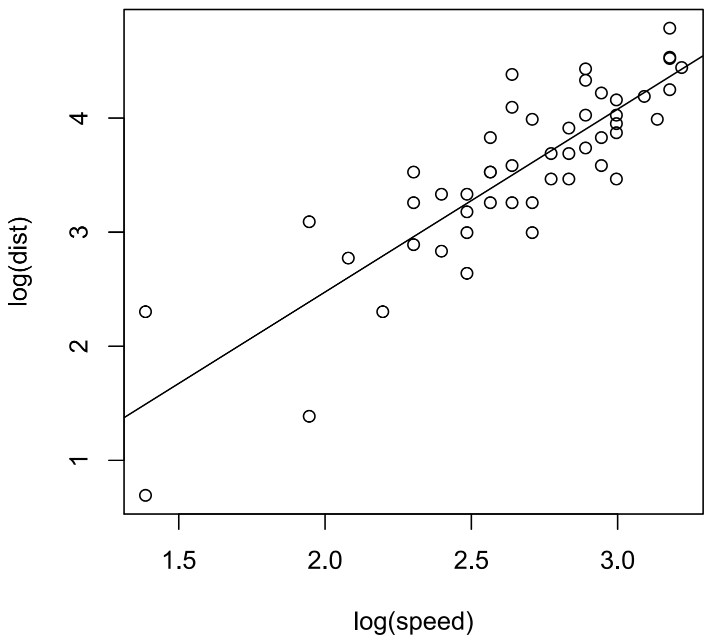
$$\sqrt{\boldsymbol{y}} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon} \qquad R^2 = 0.709$$



Response: y = distance
Predictor: x = speed

Modelling Non-Linear Relationships

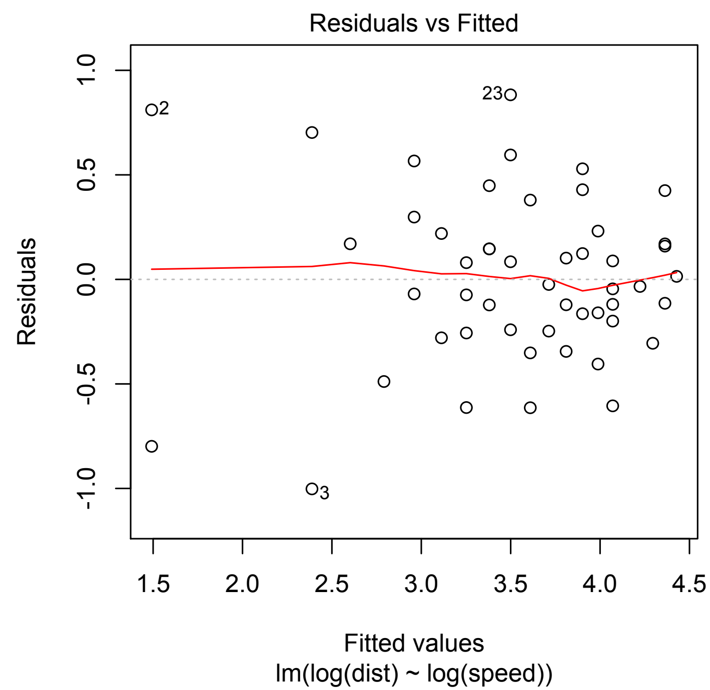
Example: *Stopping distance of cars versus speed (mph)*



$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon} \qquad R^2 = 0.651$$

$$\sqrt{\boldsymbol{y}} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon} \qquad R^2 = 0.709$$

$$\log(\boldsymbol{y}) = \alpha + \beta\log(\boldsymbol{x}) + \boldsymbol{\varepsilon} \qquad R^2 = 0.733$$

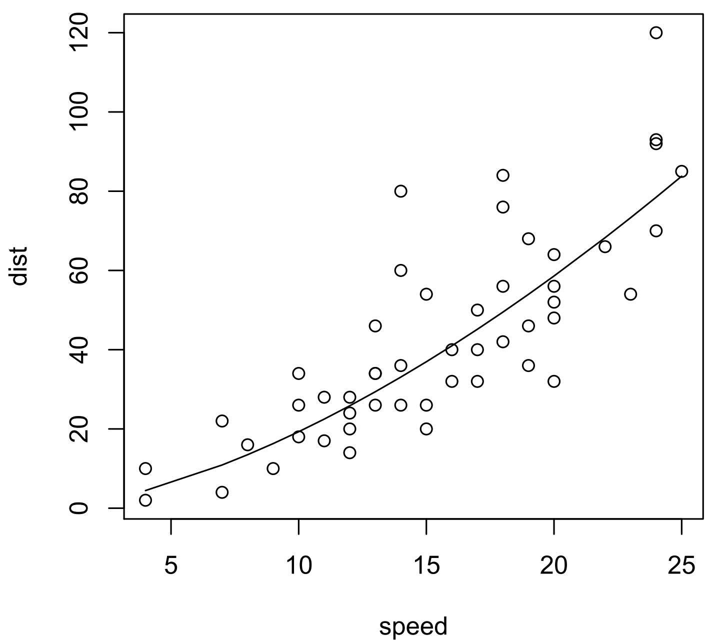


Response: y = distance

Predictor: x = speed

Modelling Non-Linear Relationships

Example: *Stopping distance of cars versus speed (mph)*



$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon} \qquad R^2 = 0.651$$

$$\sqrt{\boldsymbol{y}} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\varepsilon} \qquad R^2 = 0.709$$

$$\log(\boldsymbol{y}) = \alpha + \beta\log(\boldsymbol{x}) + \boldsymbol{\varepsilon} \qquad R^2 = 0.733$$

Response: y = distance
Predictor: x = speed

```
Call:
lm(formula = log(dist) ~ log(speed), data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00215 -0.24578 -0.02898  0.20717  0.88289

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7297     0.3758  -1.941   0.0581 .
log(speed)    1.6024     0.1395  11.484 2.26e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 0.4053 on 48 degrees of freedom
Multiple R-squared:  0.7331,  Adjusted R-squared:  0.7276
F-statistic: 131.9 on 1 and 48 DF,  p-value: 2.259e-15
```

Modelling Non-Linear Relationships

Can you use simple regression to fit this model?

$$\boldsymbol{y} = \alpha \boldsymbol{x}^{\beta} \boldsymbol{\varepsilon}$$

Non-linear
Multiplicative error model

Modelling Non-Linear Relationships

Can you use simple regression to fit this model?

$$\boldsymbol{y} = \alpha \boldsymbol{x}^{\beta} \boldsymbol{\varepsilon}$$

Non-linear
Multiplicative error model

$$\log(\boldsymbol{y}) = \log(\alpha) + \beta \log(\boldsymbol{x}) + \log(\boldsymbol{\varepsilon})$$

Modelling Non-Linear Relationships

Can you use simple regression to fit this model?

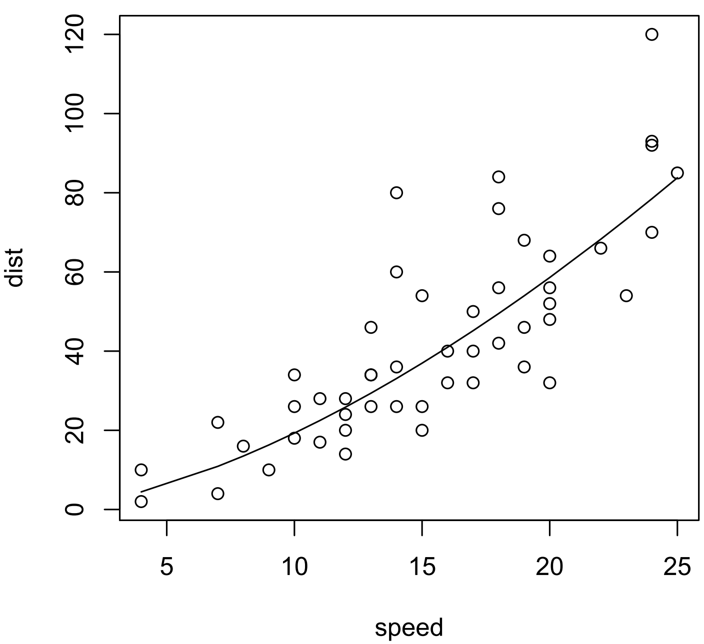
$$\boldsymbol{y} = \alpha \boldsymbol{x}^{\beta} \boldsymbol{\varepsilon}$$

Non-linear
Multiplicative error model

$$\log(\boldsymbol{y}) = \log(\alpha) + \beta \log(\boldsymbol{x}) + \log(\boldsymbol{\varepsilon})$$

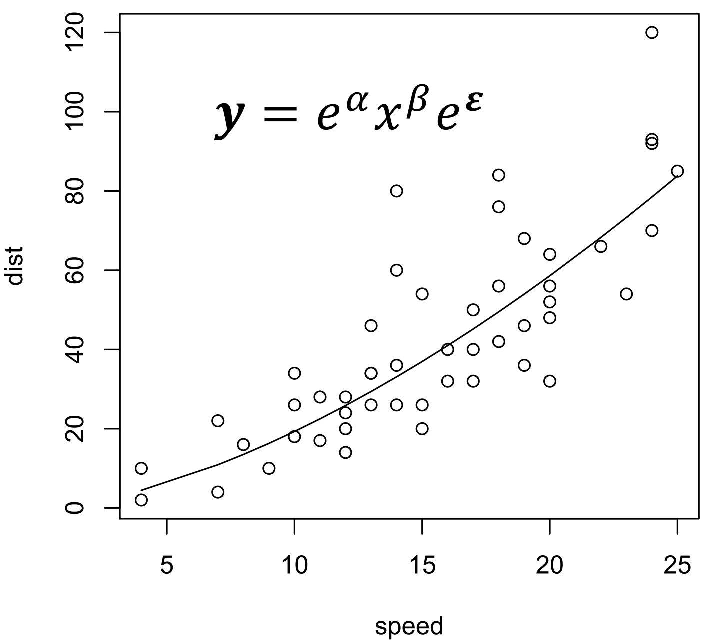
Yes, so long as $\log(\boldsymbol{\varepsilon}) \sim N(0, \sigma^2)$

Error model is log-Normal.



Modelling Non-Linear Relationships

Example: *Stopping distance of cars versus speed (mph)*



$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon} \qquad R^2 = 0.651$$

$$\sqrt{\boldsymbol{y}} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon} \qquad R^2 = 0.709$$

$$\log(\boldsymbol{y}) = \alpha + \beta \log(\boldsymbol{x}) + \boldsymbol{\varepsilon} \qquad R^2 = 0.733$$

```
Call:
lm(formula = log(dist) ~ log(speed), data = cars)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.00215 -0.24578 -0.02898  0.20717  0.88289 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.7297     0.3758  -1.941   0.0581 .  
log(speed)    1.6024     0.1395  11.484 2.26e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 
‘ ’ 1

Residual standard error: 0.4053 on 48 degrees of freedom
Multiple R-squared:  0.7331,	Adjusted R-squared:  0.7276 
F-statistic: 131.9 on 1 and 48 DF,  p-value: 2.259e-15
```

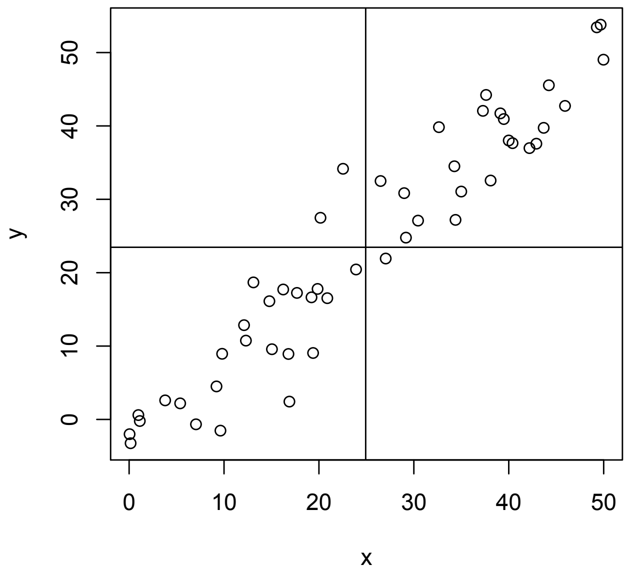
Response: y = distance
Predictor: x = speed

Simple Regression in R:

Correlation Coefficients:

R functions:

plot(x,y)
cor(x,y)
cor.test(x,y)



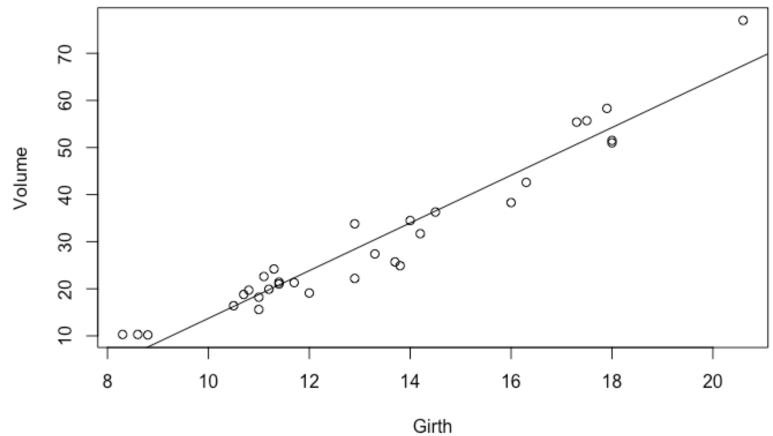
data: x and y
t = 17.613, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8802556 0.9602168
sample estimates:
cor
0.9305923

Simple Regression in R:

R functions:

plot(x,y)
m1 <- lm(y~x)
abline(m1)

summary(m1)

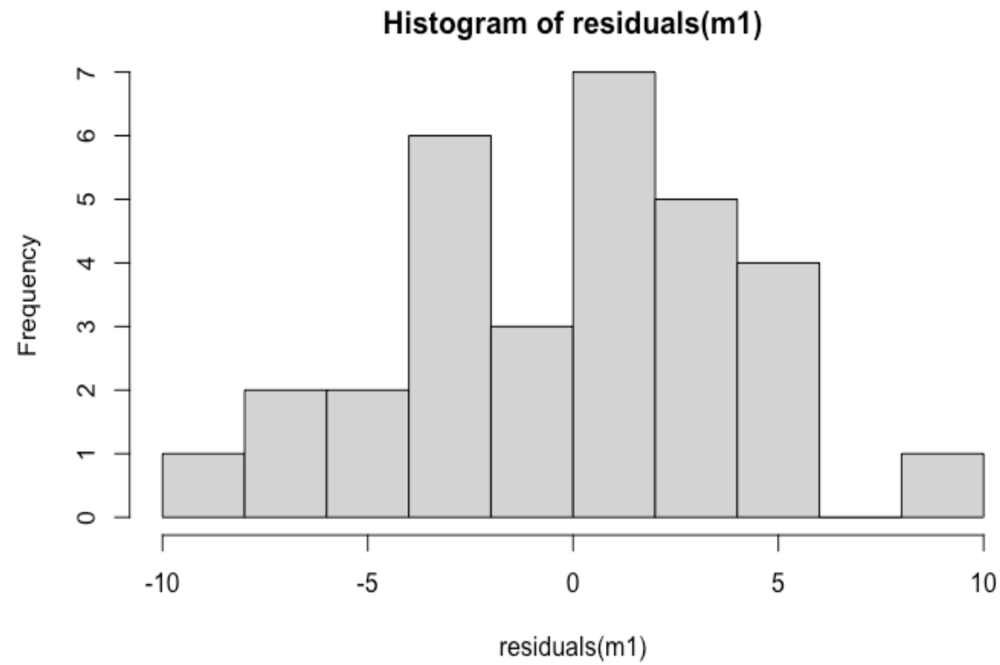



```
Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min     1Q Median     3Q    Max
-8.065 -3.107  0.152  3.495  9.587

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,	Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Simple Regression in R:

R functions:

```
plot(x,y)
m1 <- lm(y~x)
abline(m1)

summary(m1)

r1 <- residuals(m1)
hist(r1)
```



Simple Regression in R:

R functions:

plot(x,y)
m1 <- lm(y~x)
abline(m1)

summary(m1)

r1 <- residuals(r1)
hist(r1)

plot(m1)