



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE



UNIVERSITY OF
CAMBRIDGE

Linear Modelling: Simple Regression

Cancer Research UK – 19th of July 2017

D.-L. Couturier / M. Dunning / R. Nicholls

Introduction:

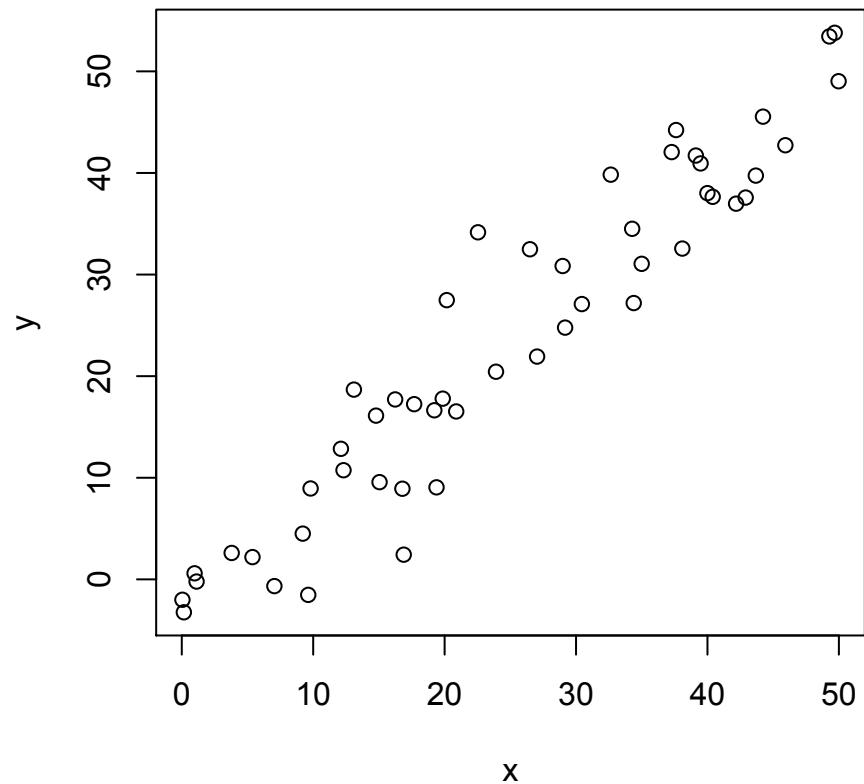
ANOVA

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

Regression

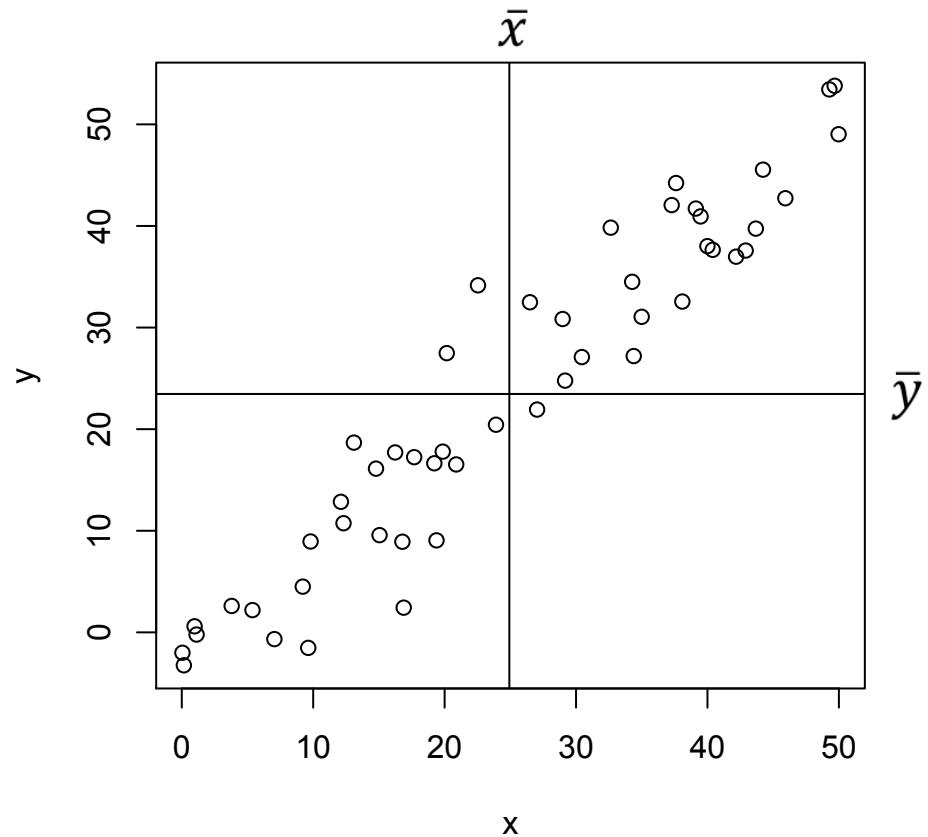
- Used for revealing and investigating relationships between input and output variables
- Model data, and extrapolate as much information as possible

Correlation:

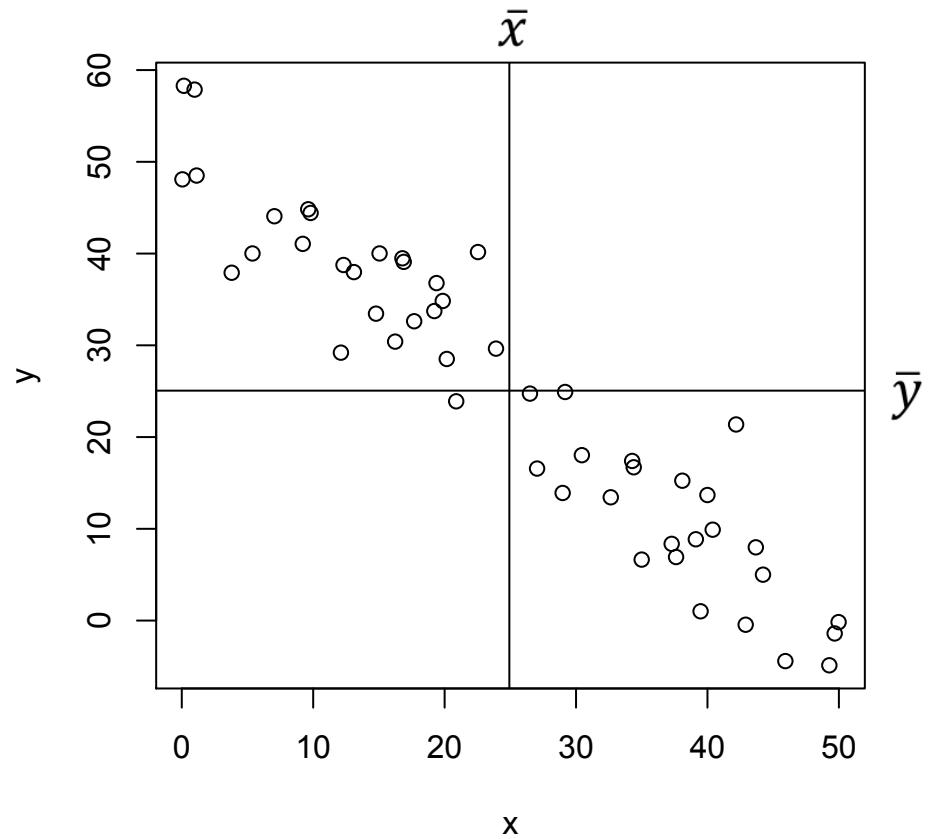


How to measure the strength of a linear relationship between variables?

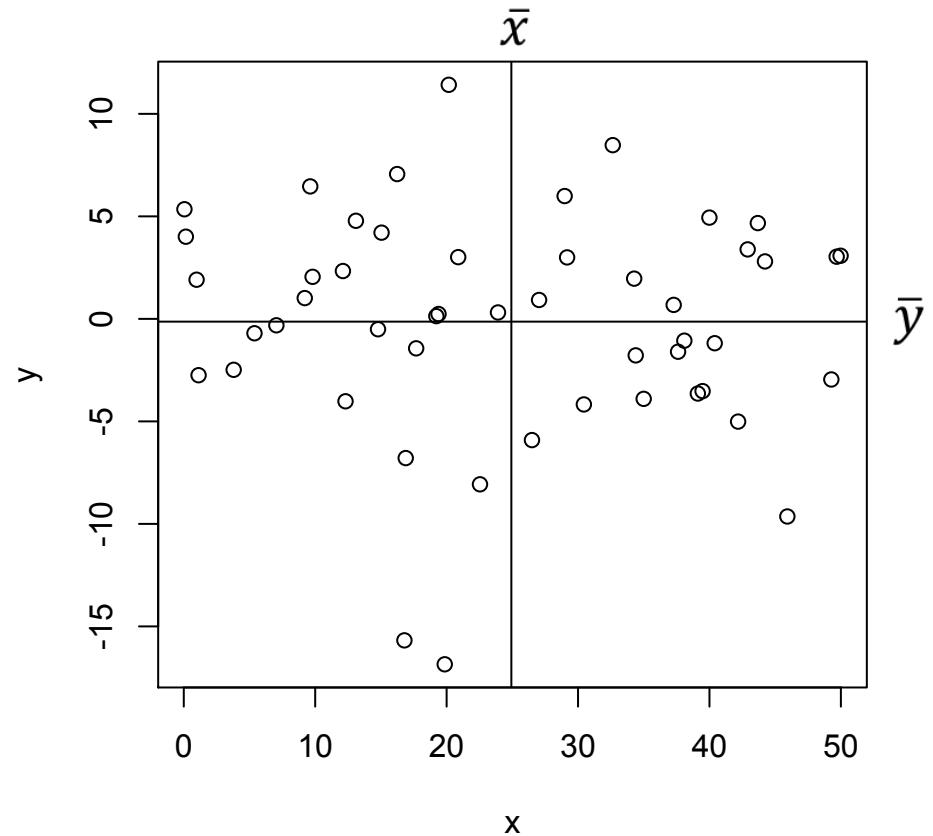
Correlation:



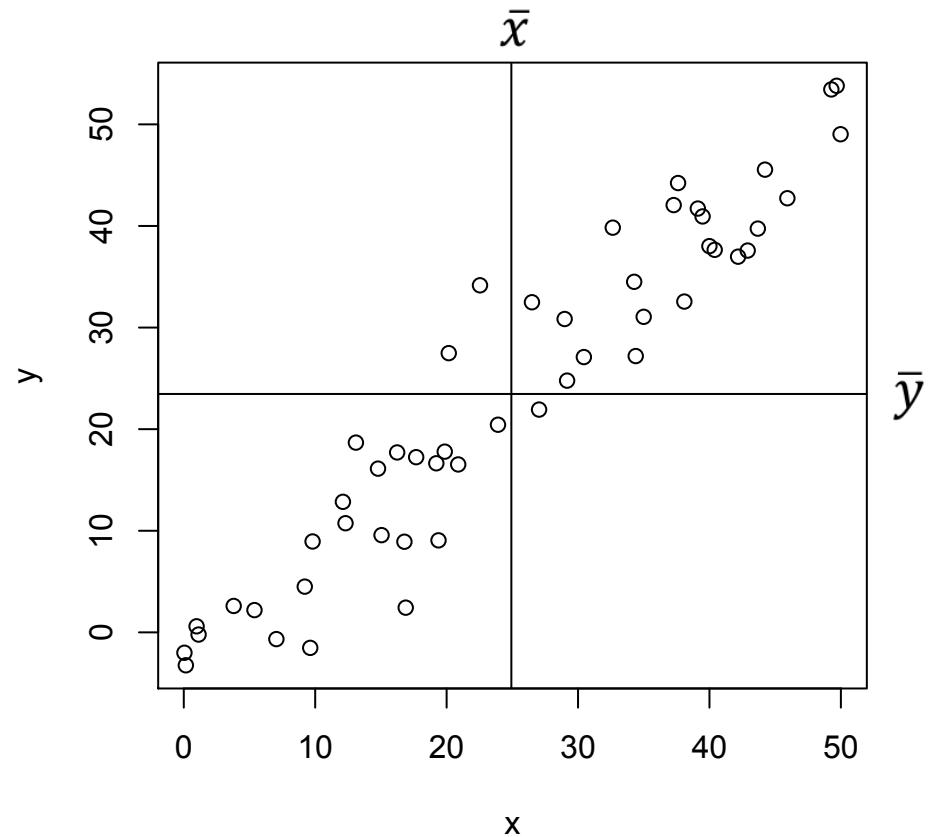
Correlation:



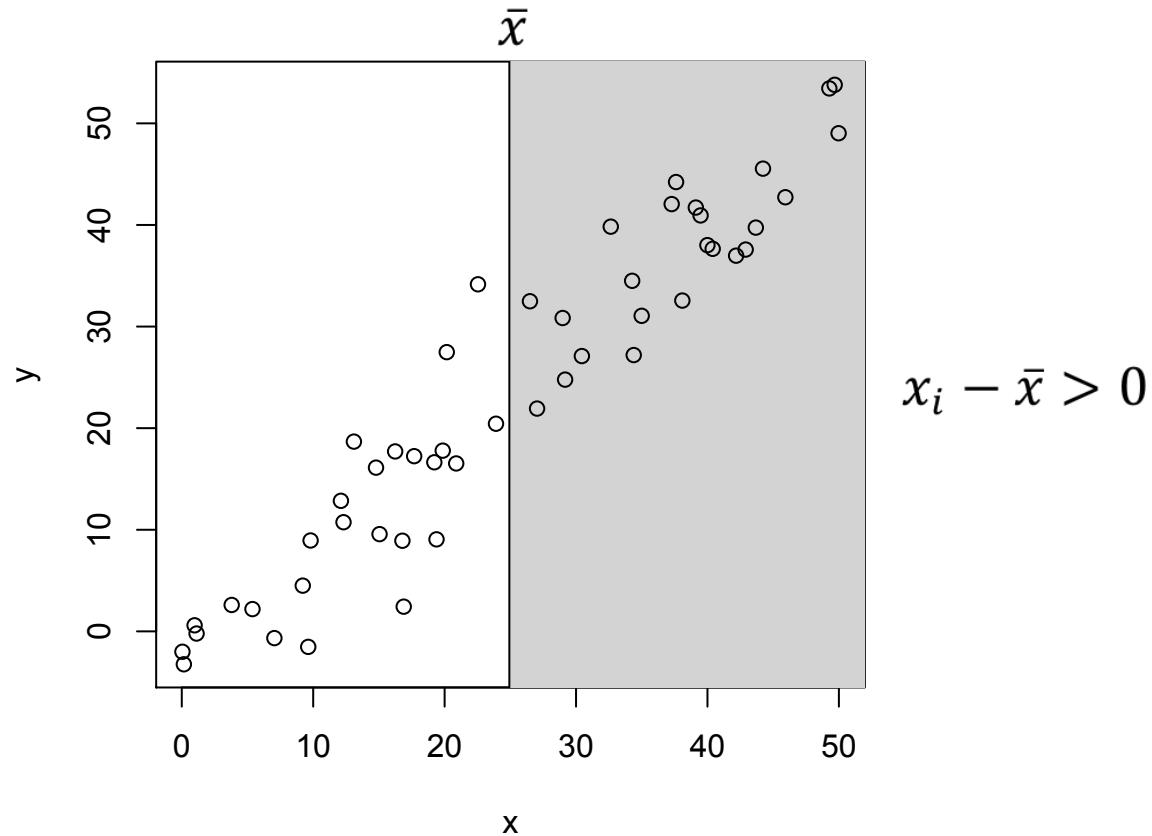
Correlation:



Correlation:

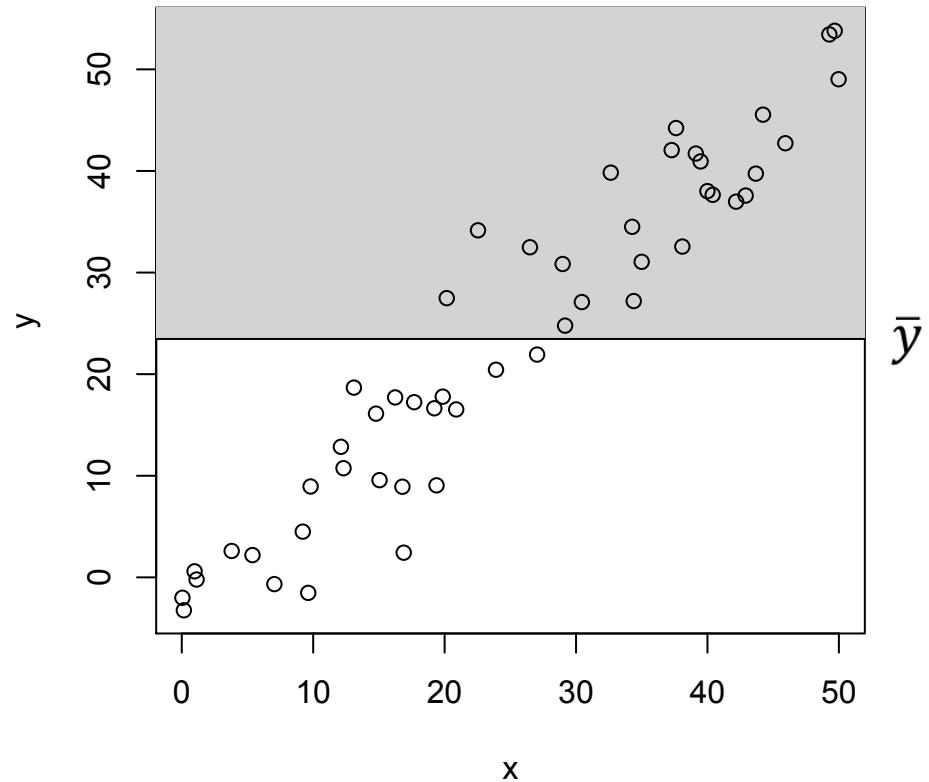


Correlation:

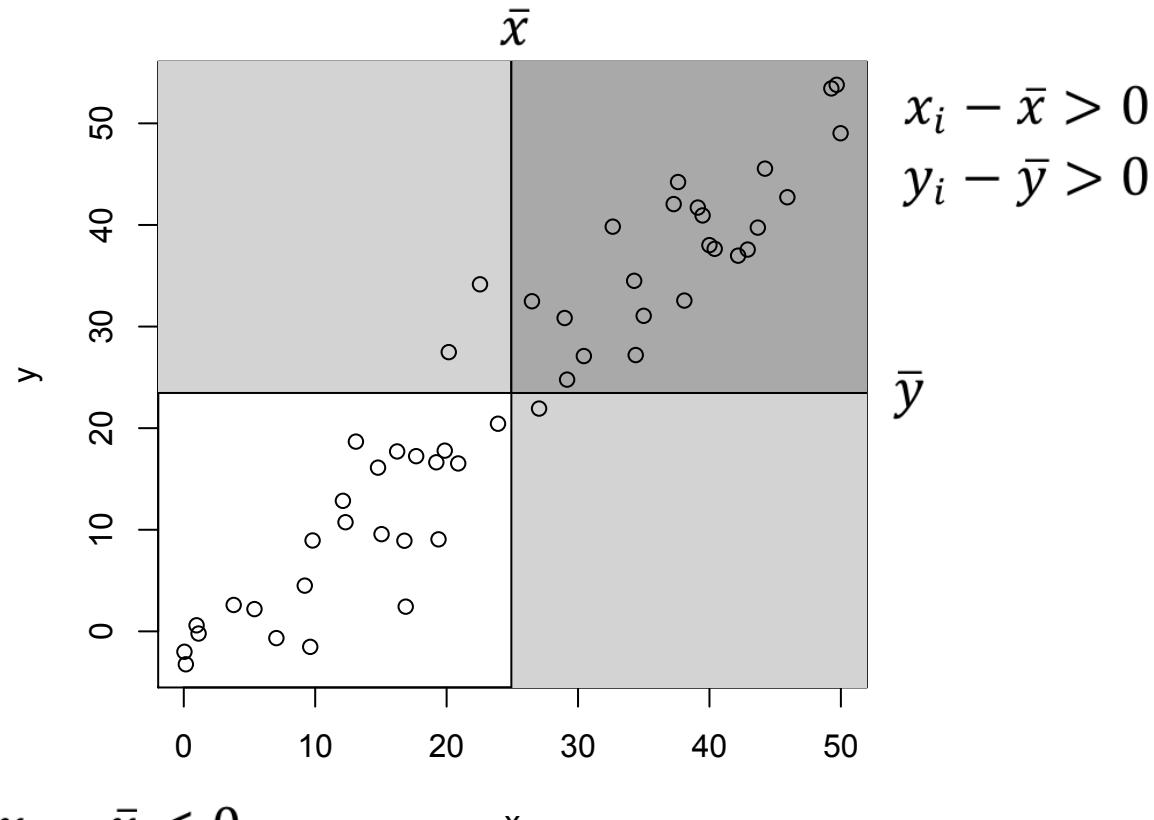


Correlation:

$$y_i - \bar{y} > 0$$



Correlation:



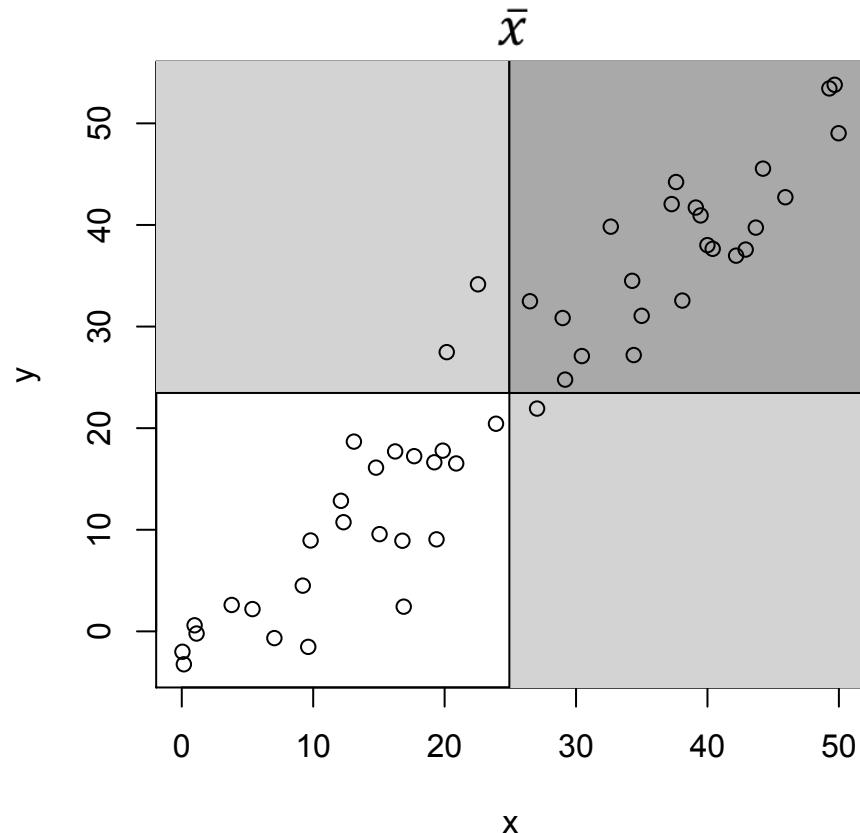
$$x_i - \bar{x} < 0$$

$$y_i - \bar{y} < 0$$

Correlation:

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

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



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

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

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:

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

Correlation:

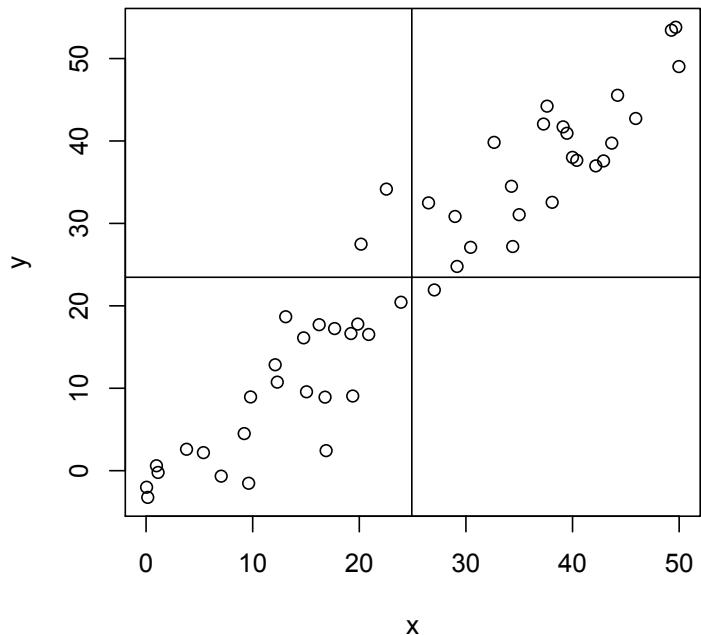
Pearson's product-moment correlation coefficient:

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

Coefficient of Variation (R^2 value):

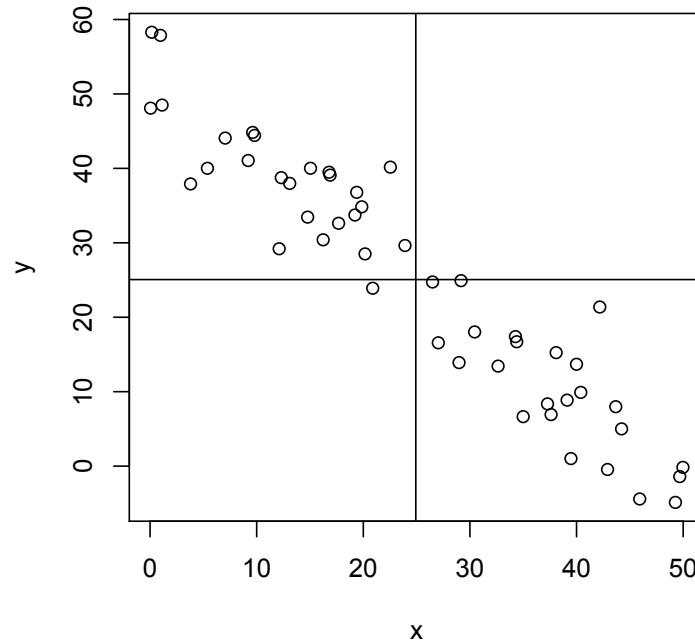
$$R_{X,Y}^2 = r_{X,Y}^2$$

Correlation:



$$r = 0.931$$

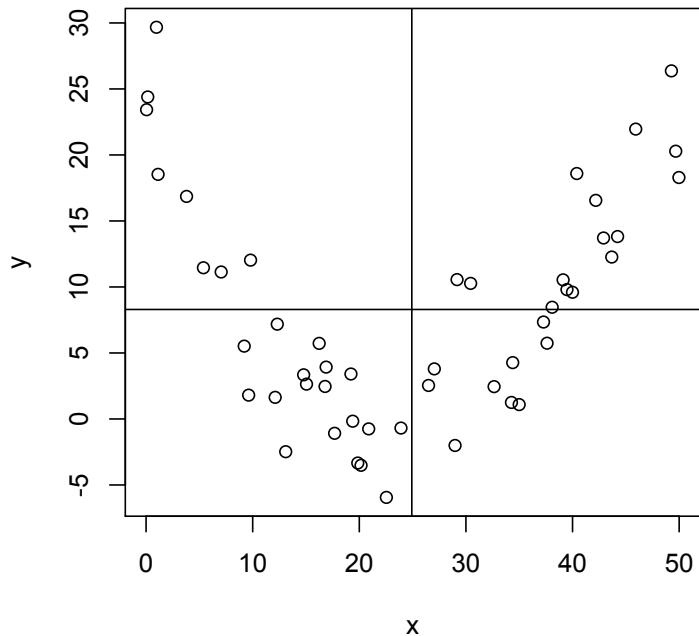
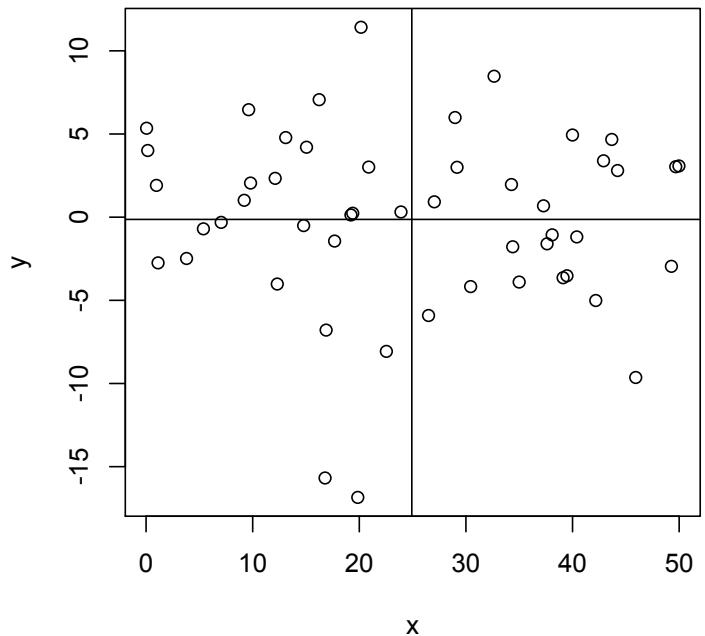
$$R^2 = 0.866$$



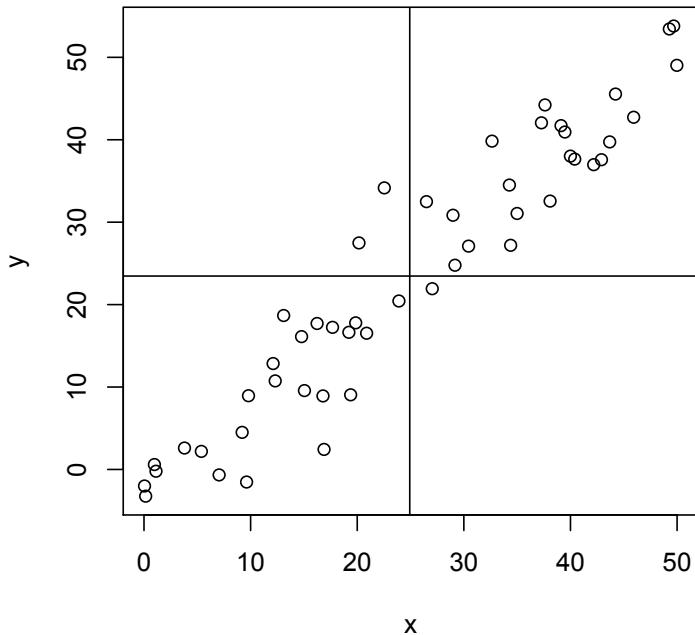
$$r = -0.949$$

$$R^2 = 0.901$$

Correlation:



Correlation:



data: x and y
t = 17.613, df = 48, p-value < 2.2e-16

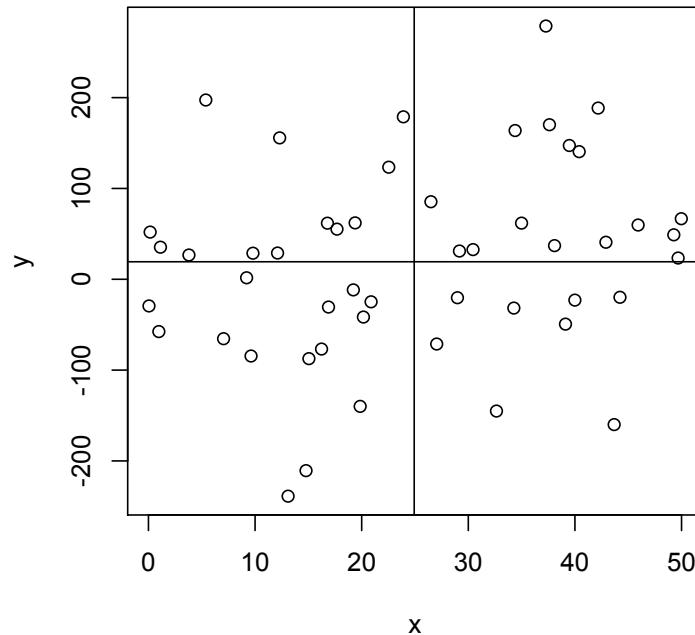
alternative hypothesis: true correlation is not equal to 0

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:

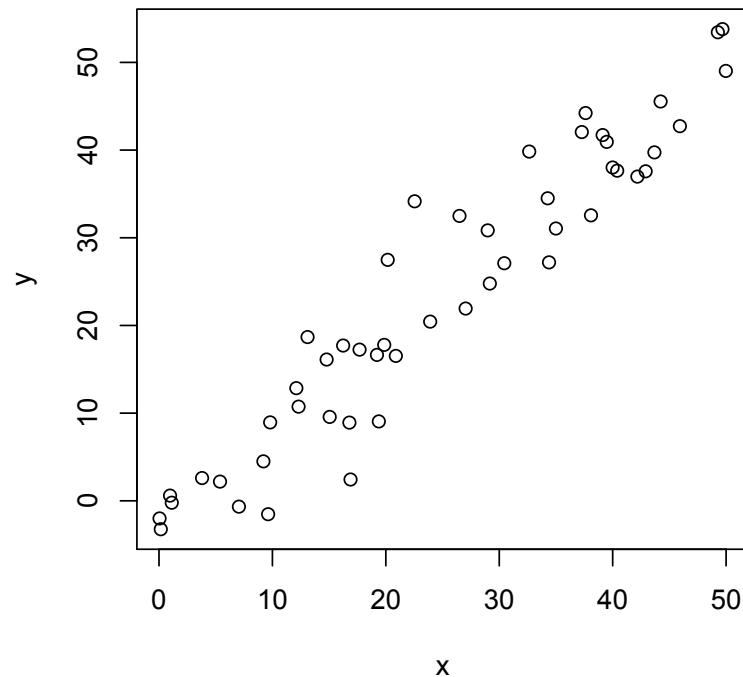
cor
0.2197833

Simple Regression:

Aims:

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

Response variable
Dependent variable



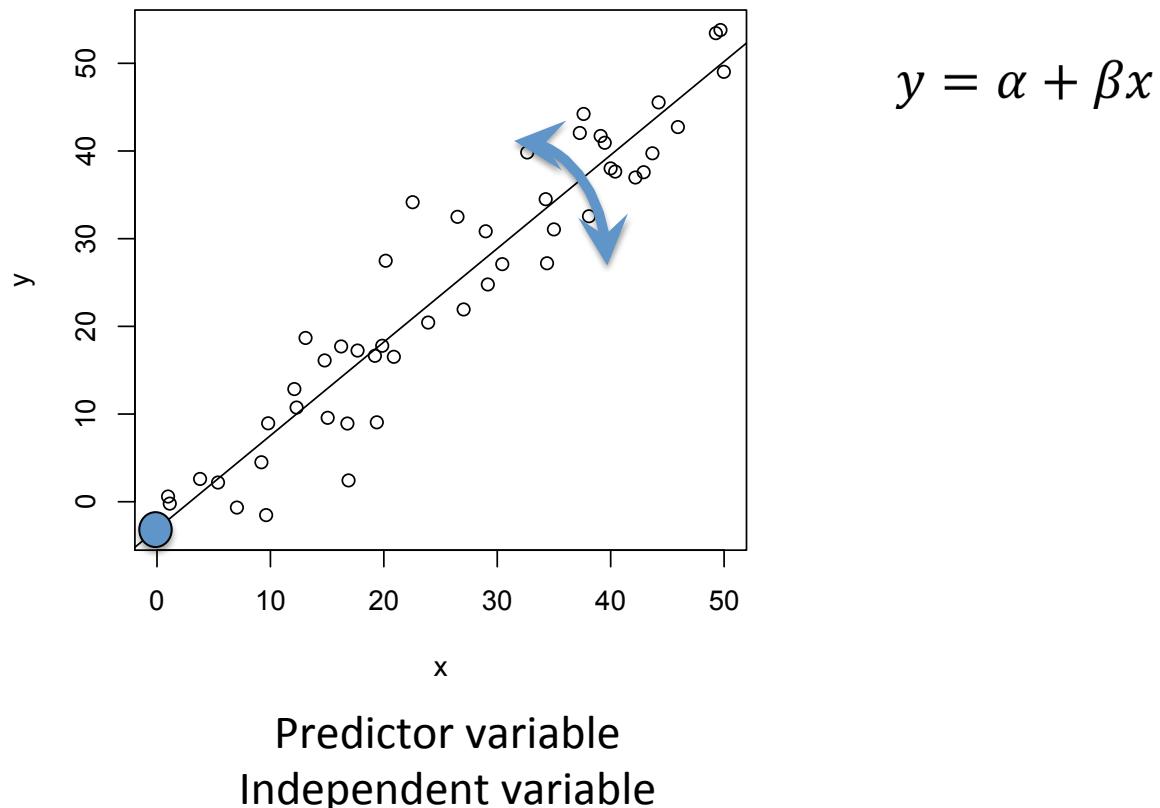
Predictor variable
Independent variable

Simple Regression:

Aims:

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

Response variable
Dependent variable

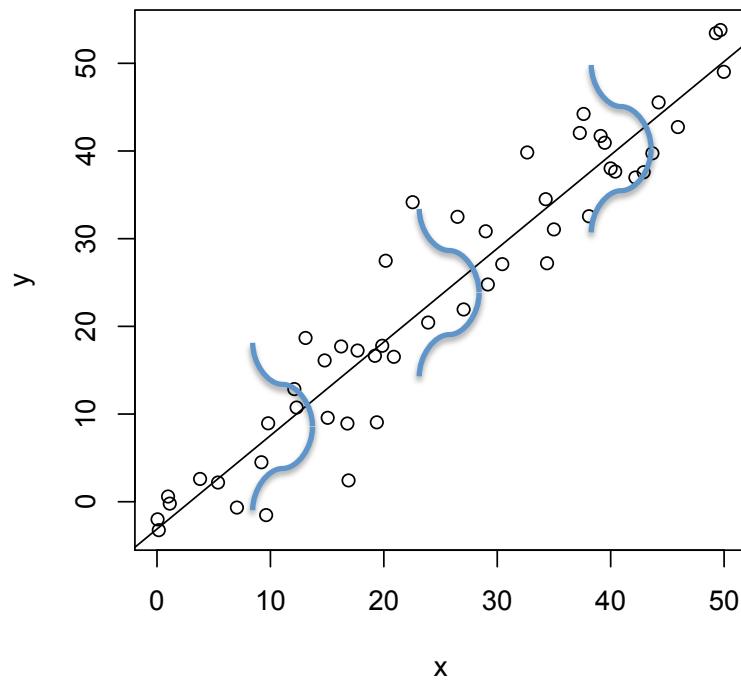


Simple Regression:

Aims:

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

Response variable
Dependent variable



Predictor variable
Independent variable

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

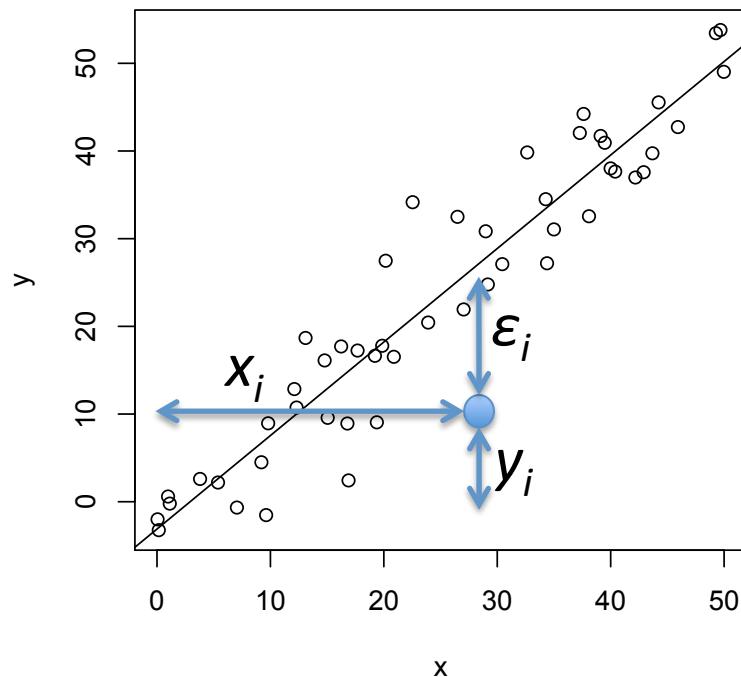
$$\varepsilon \sim N(0, \sigma^2)$$

Simple Regression:

Aims:

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

Response variable
Dependent variable



Predictor variable
Independent variable

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

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

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

ε_i = errors, residuals

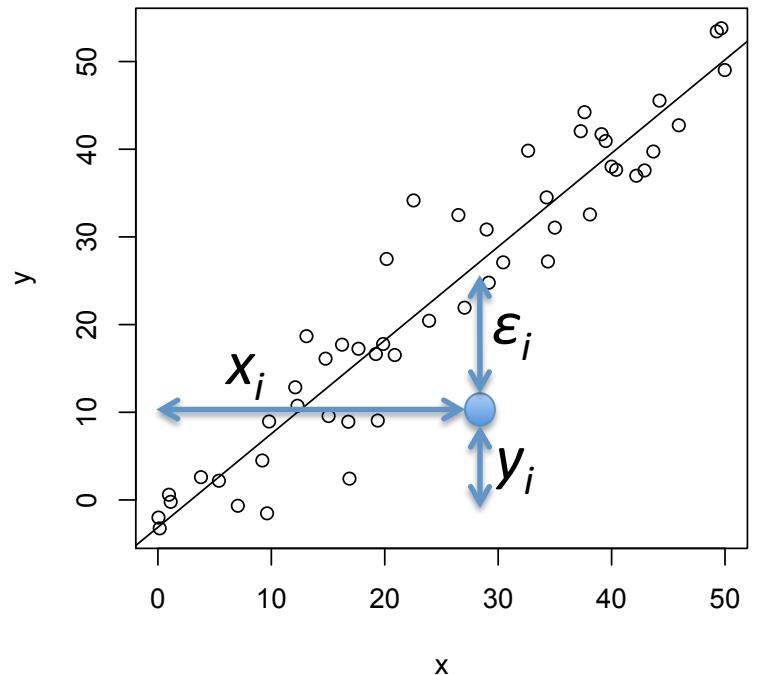
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

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

Simple Regression:

So how do we fit the regression line?

Find optimal parameter estimates $\hat{\alpha}$ and $\hat{\beta}$



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

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

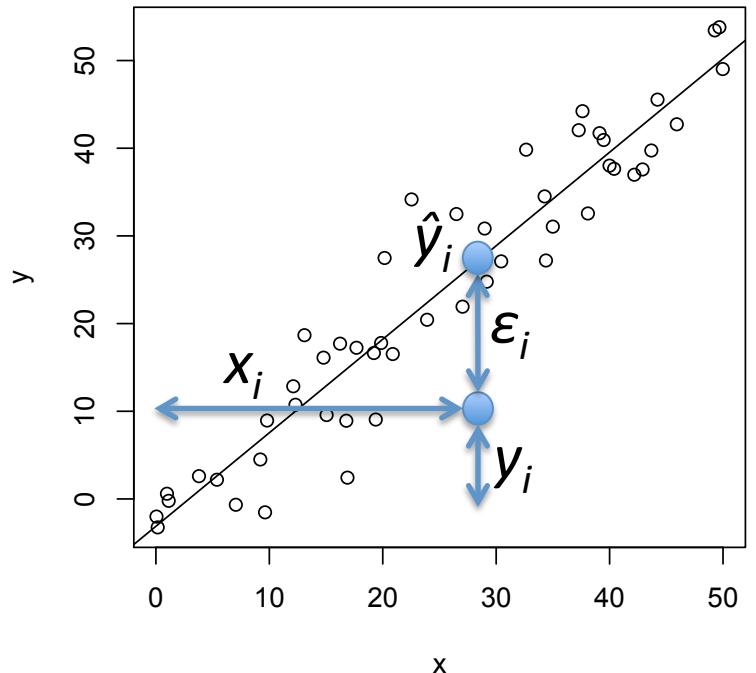
$$\mathbf{y} = \alpha + \beta \mathbf{x} + \boldsymbol{\varepsilon}$$

So how do we fit the regression line?

Find optimal parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

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

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



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

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

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

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

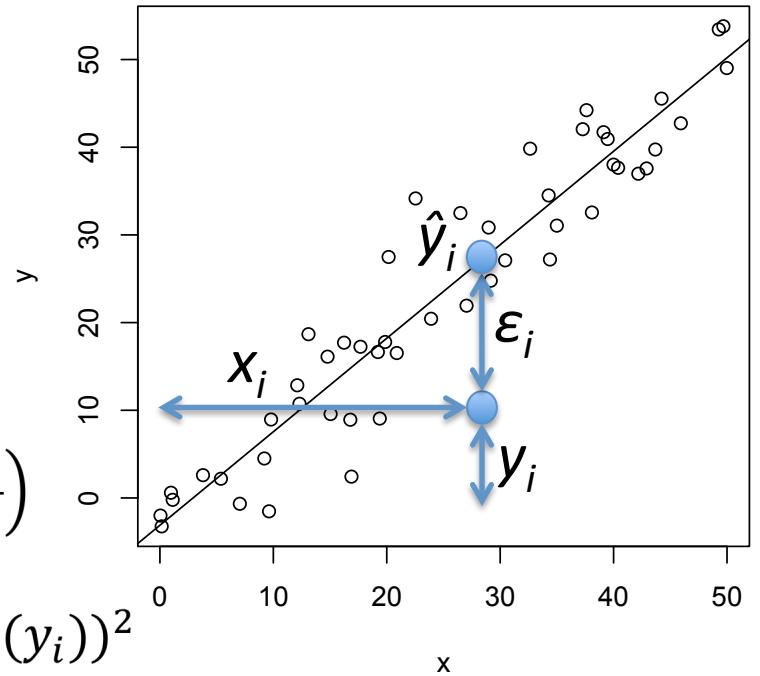
Simple Regression:

So how do we fit the regression line?

Find optimal parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

Maximise likelihood of parameters given the data:

$$\begin{aligned}\mathcal{L}(\alpha, \beta | \mathbf{y}, \mathbf{x}) &= \prod_i f(y_i | x_i; \alpha, \beta) \\ &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - E(y_i))^2}{2\sigma^2}} \\ &= \sum_i \left(\frac{-1}{2} \log(2\pi\sigma^2) - \frac{(y_i - E(y_i))^2}{2\sigma^2} \right) \\ &= \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - E(y_i))^2\end{aligned}$$



$$\mathcal{L}(\alpha, \beta | \mathbf{y}, \mathbf{x}) = \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - E(y_i))^2$$

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

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

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

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

So how do we fit the regression line?

Find optimal parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

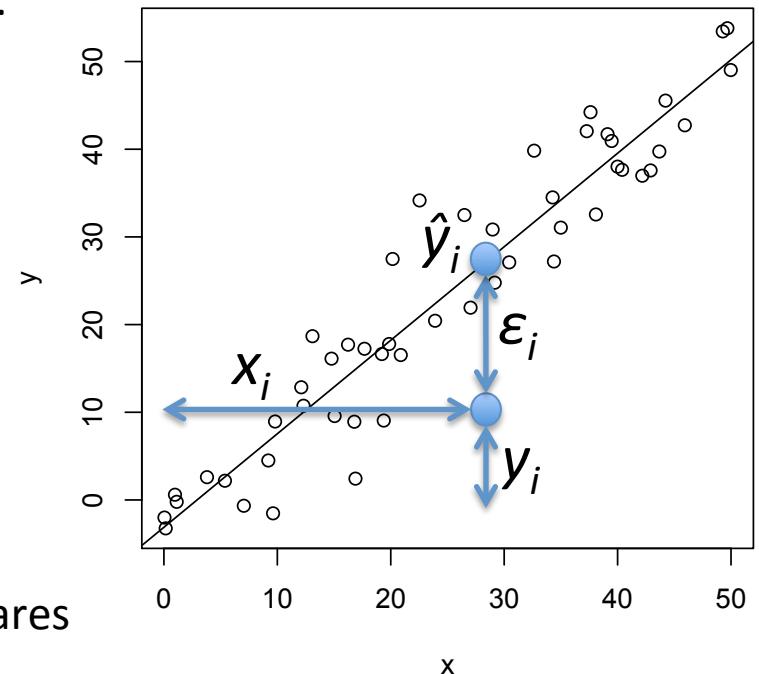
Maximise likelihood of parameters given the data:

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

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

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

Optimal parameters : minimise residual sum of squares



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

$$\mathcal{L}(\alpha, \beta | \mathbf{y}, \mathbf{x}) = \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - E(y_i))^2$$

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

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

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

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

So how do we fit the regression line?

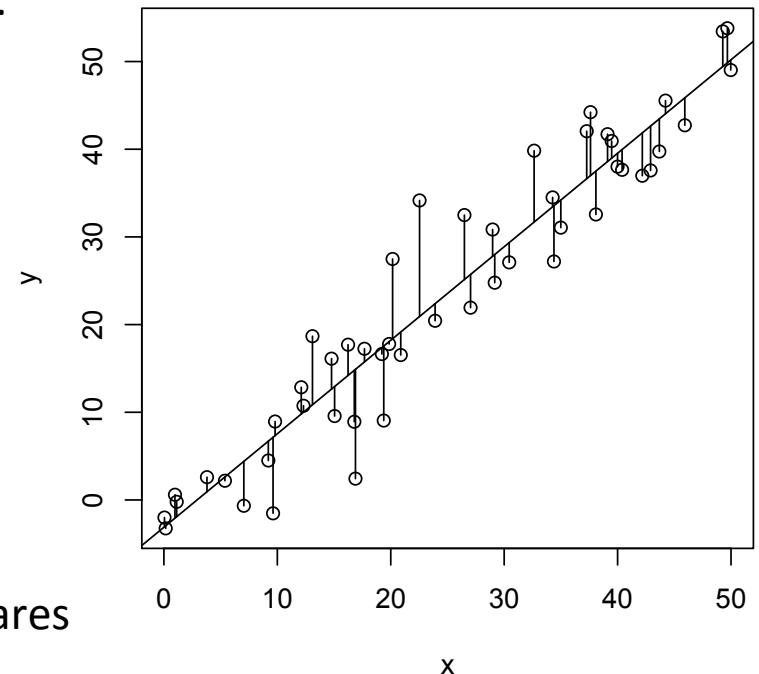
Find optimal parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

Maximise likelihood of parameters given the data:

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

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

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



Optimal parameters : minimise residual sum of squares

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

$$\mathcal{L}(\alpha, \beta | \mathbf{y}, \mathbf{x}) = \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - E(y_i))^2$$

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

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

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

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

So how do we fit the regression line?

Find optimal parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

Maximise likelihood of parameters given the data:

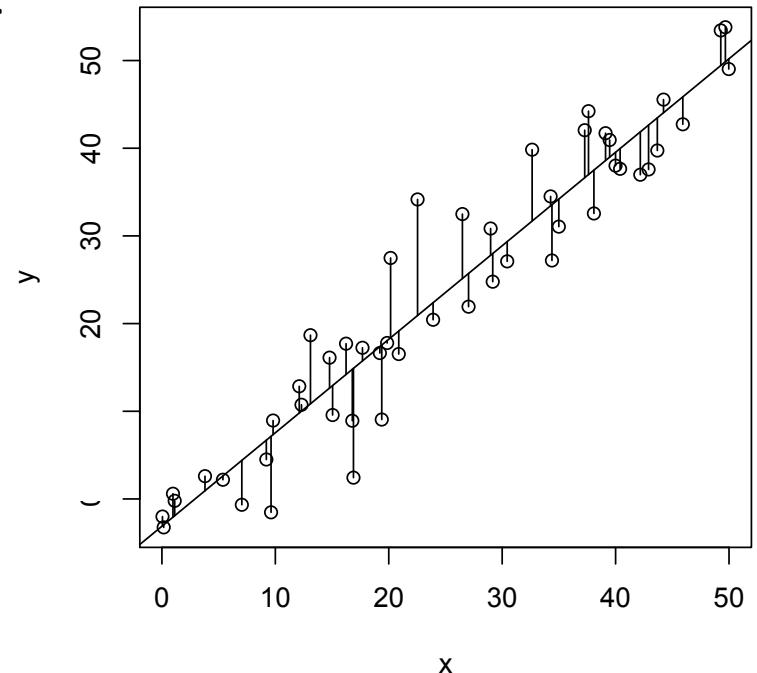
$$\mathcal{L}(\alpha, \beta | \mathbf{y}, \mathbf{x}) \rightarrow \max$$

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

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

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

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



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:

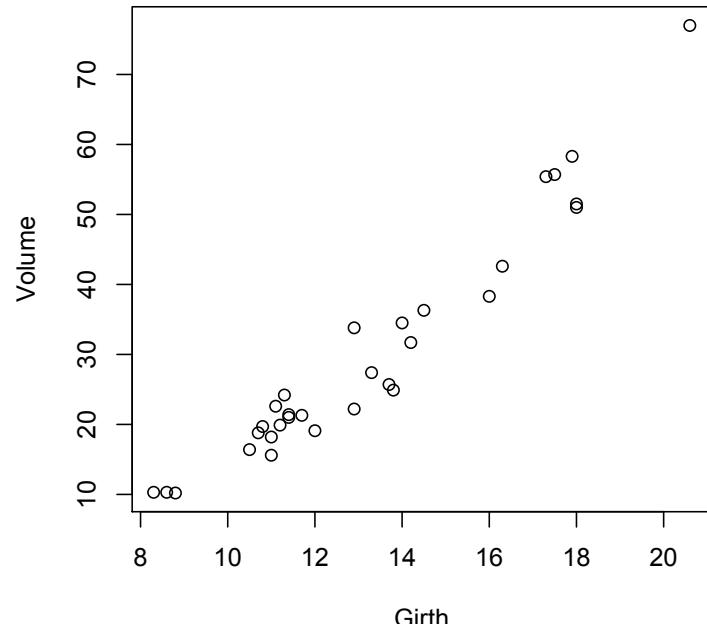
Min	1Q	Median	3Q	Max
-8.065	-3.107	0.152	3.495	9.587

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: $y = \text{Volume}$
Predictor: $x = \text{Girth}$

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:

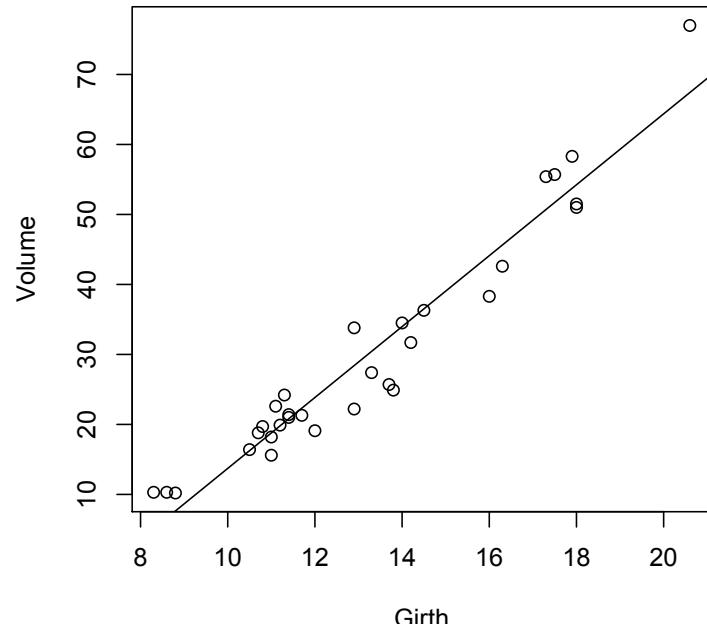
Min	1Q	Median	3Q	Max
-8.065	-3.107	0.152	3.495	9.587

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: $y = \text{Volume}$
Predictor: $x = \text{Girth}$

$$y = -36.9 + 5.07x$$

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:

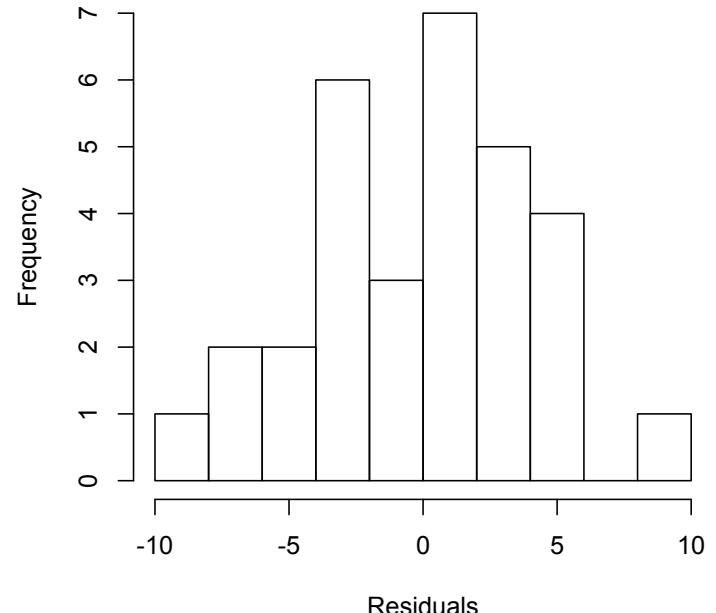
Min	1Q	Median	3Q	Max
-8.065	-3.107	0.152	3.495	9.587

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



Response: $y = \text{Volume}$
Predictor: $x = \text{Girth}$

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

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:

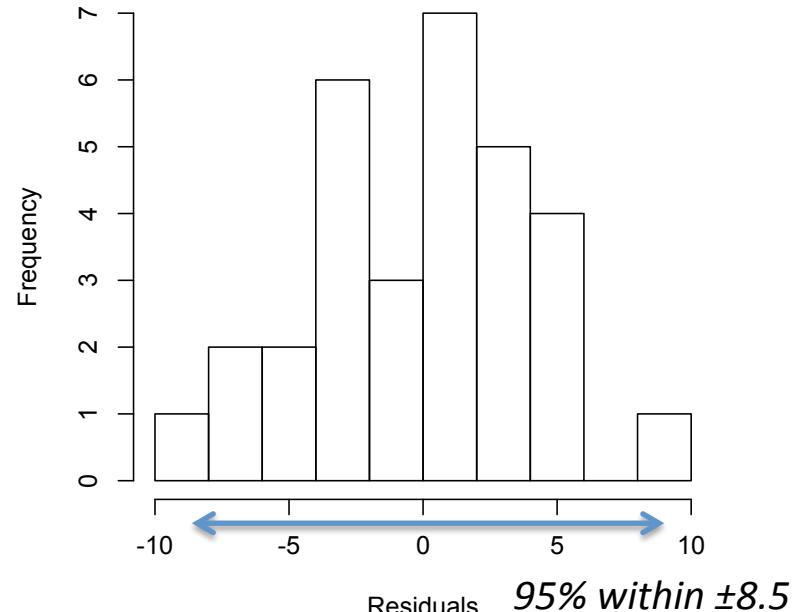
Min	1Q	Median	3Q	Max
-8.065	-3.107	0.152	3.495	9.587

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



Response: $y = \text{Volume}$
Predictor: $x = \text{Girth}$

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

Linear Regression:

Assumptions:

1. Model is linear in parameters.

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

Linear Regression:

Assumptions:

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

$$y = \alpha + \beta x^2 + \varepsilon$$

Linear Regression:

Assumptions:

1. Model is linear in parameters.

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

$$y = \alpha + \beta x^2 + \varepsilon$$

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

Linear Regression:

Assumptions:

1. Model is linear in parameters.

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

$$y = \alpha + \beta x^2 + \varepsilon$$

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

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

Linear Regression:

Assumptions:

1. Model is linear in parameters.

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

$$y = \alpha + \beta x^2 + \varepsilon$$

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

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

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

Linear Regression:

Assumptions:

1. Model is linear in parameters.
2. Gaussian error

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

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Linear Regression:

Assumptions:

1. Model is linear in parameters.
2. Gaussian error
3. Additive error model.

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

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

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

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

Linear Regression:

Assumptions:

1. Model is linear in parameters.

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

2. Gaussian error

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

3. Additive error model.

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

4. Independence of errors.

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$$

No autocorrelation – when one observation depends on the last

Linear Regression:

Assumptions:

1. Model is linear in parameters.

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

2. Gaussian error model.

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

3. Additive error model.

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

4. Independence of errors.

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$$

No autocorrelation – when one observation depends on the last

5. Homoscedasticity.

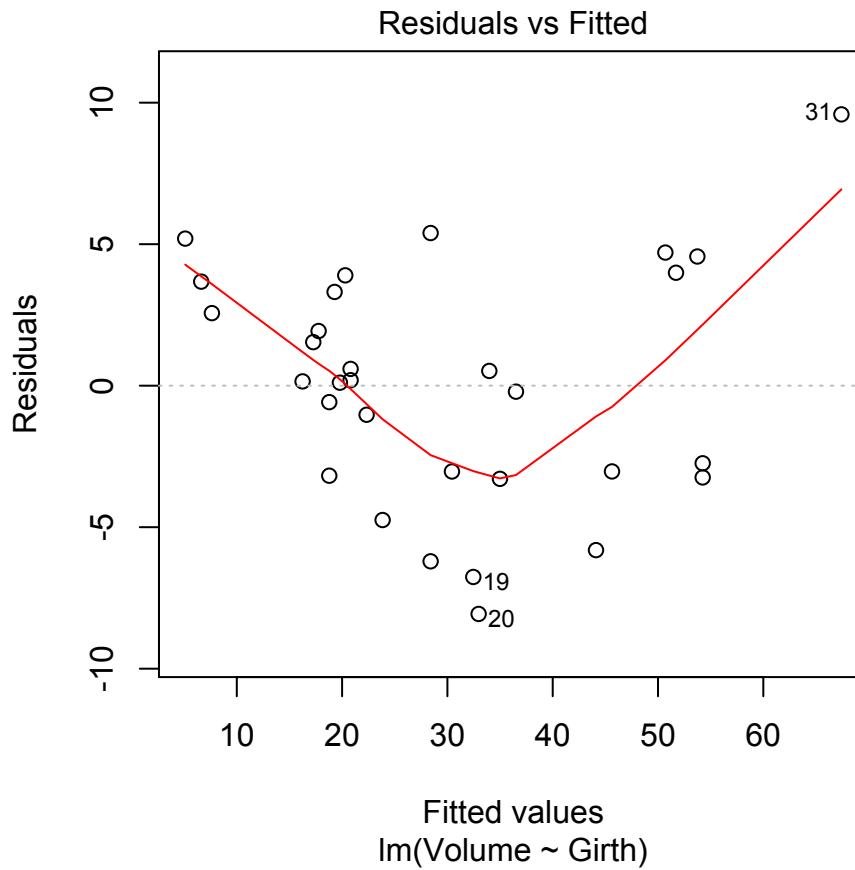
$$\text{Var}(\varepsilon|x) = \sigma^2 \mathbf{I}$$

Homogeneity / stability of variance of the residuals

Testing Assumptions: diagnostic plots

1. Residuals vs Fitted Values

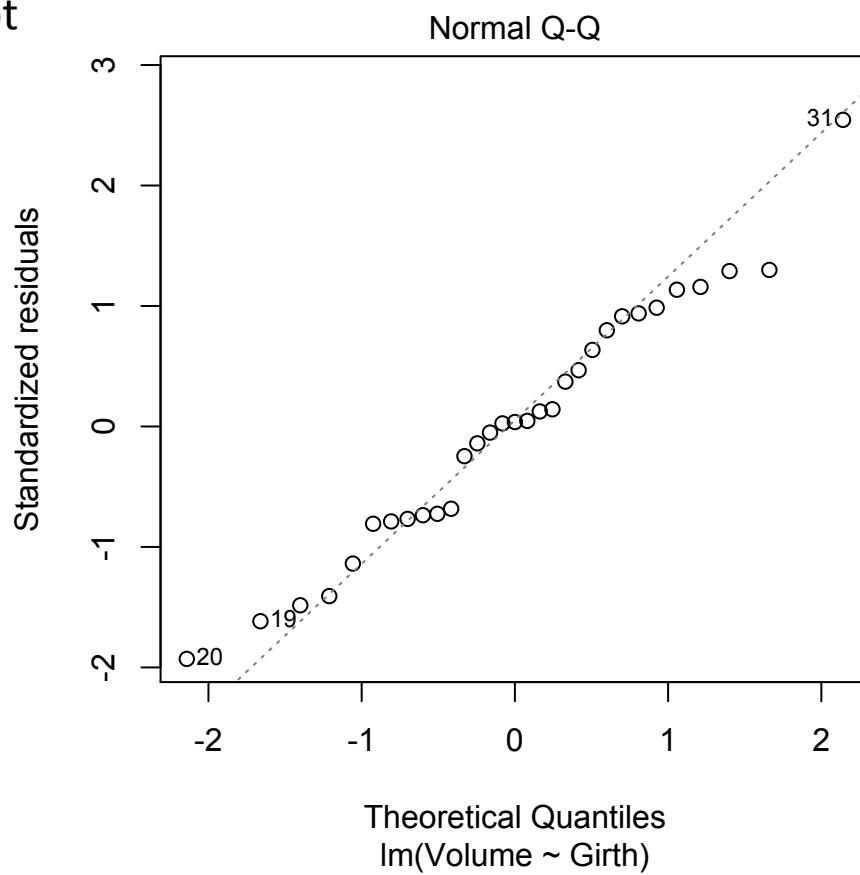
- Should not be related
- No visible pattern
- Mean residual = zero
- Constant variance



Testing Assumptions: diagnostic plots

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

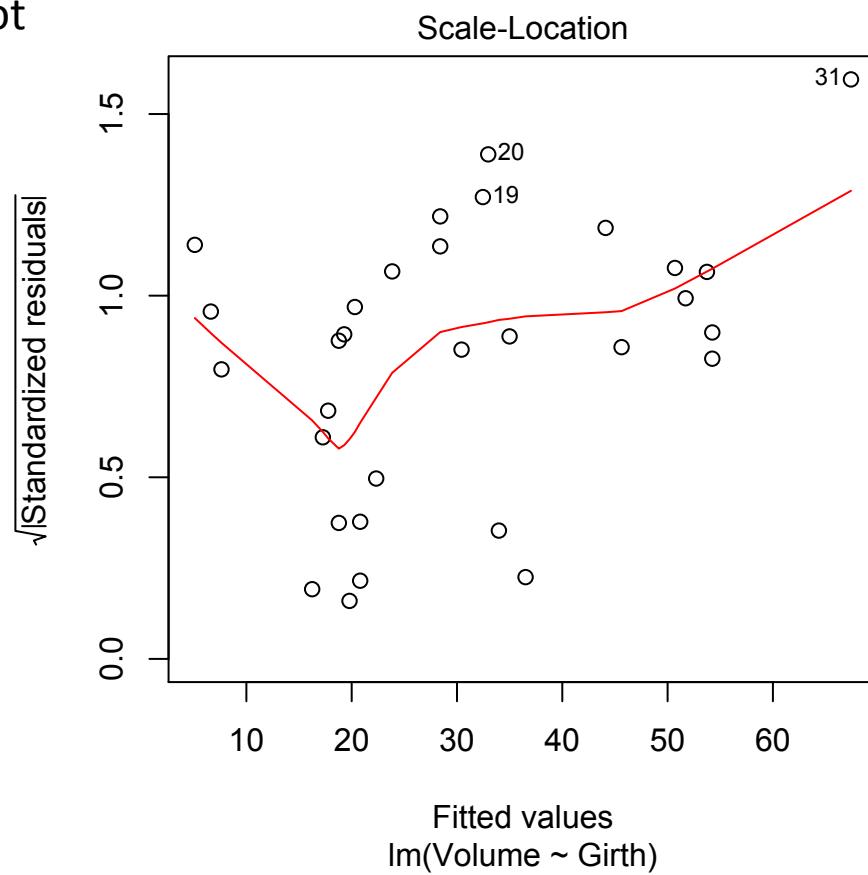
- Visual test for Normality
- No strong trends/departures



Testing Assumptions: diagnostic plots

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

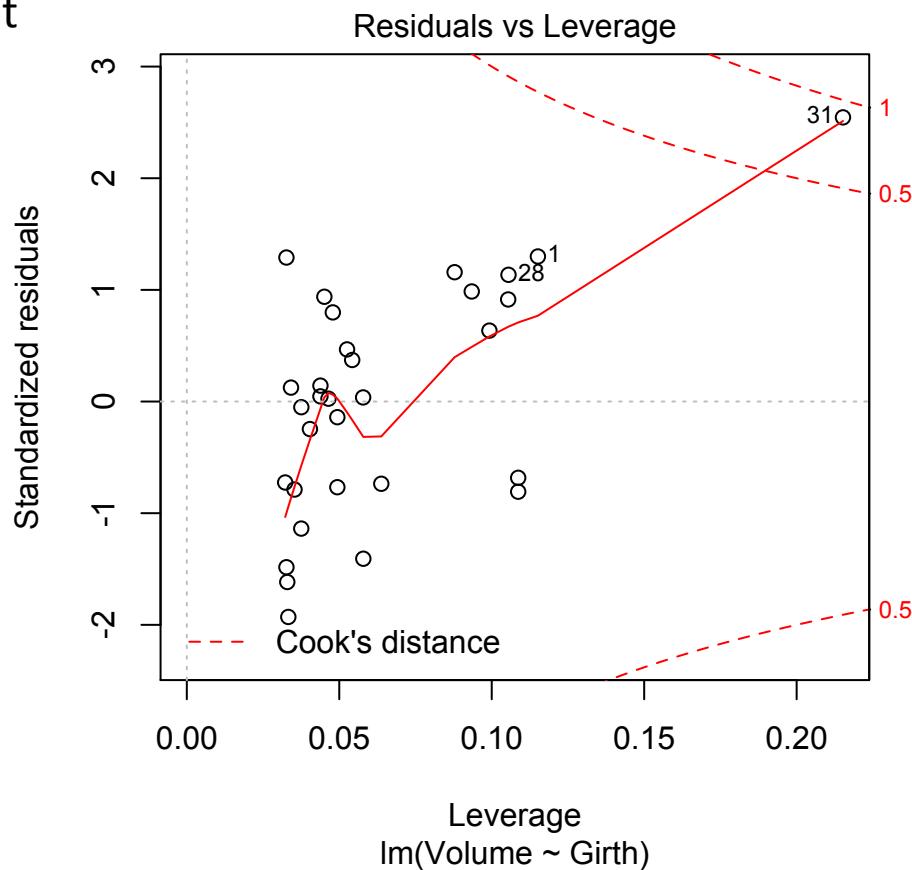
- Test for homoscedasticity
- Should be constant, ≈ 1
- No trend



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...

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

$$y = \alpha + \beta x^2 + \varepsilon$$

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

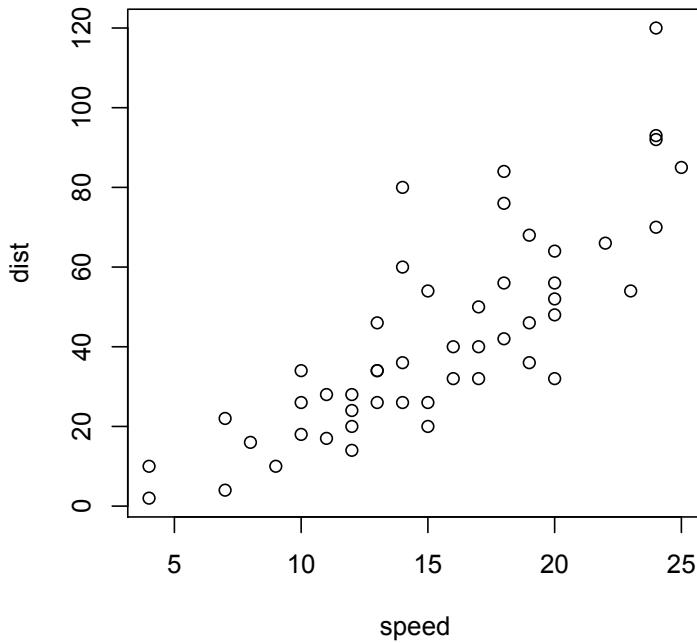
$$\log(y) = \alpha + \beta \sqrt{x} + \varepsilon$$

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.

Modelling Non-Linear Relationships

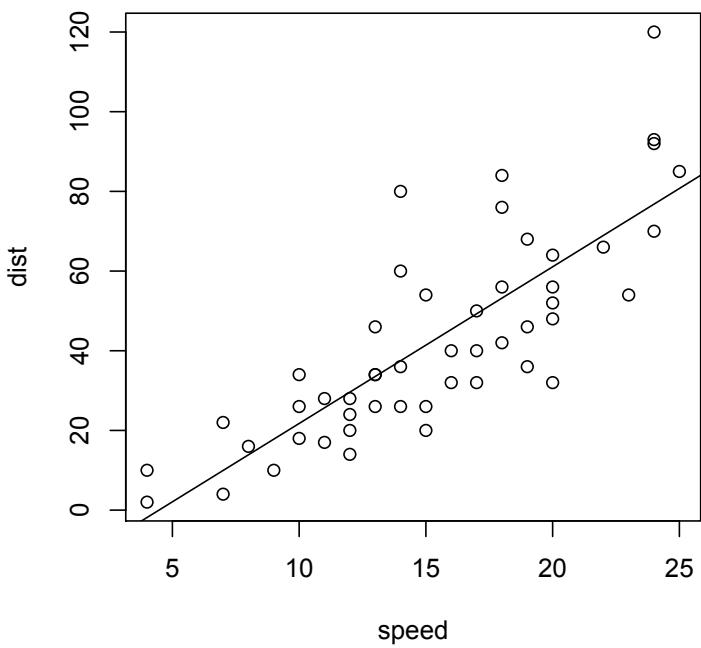
Example: *Stopping distance of cars versus speed (mph)*



Response: $y = \text{distance}$
Predictor: $x = \text{speed}$

Modelling Non-Linear Relationships

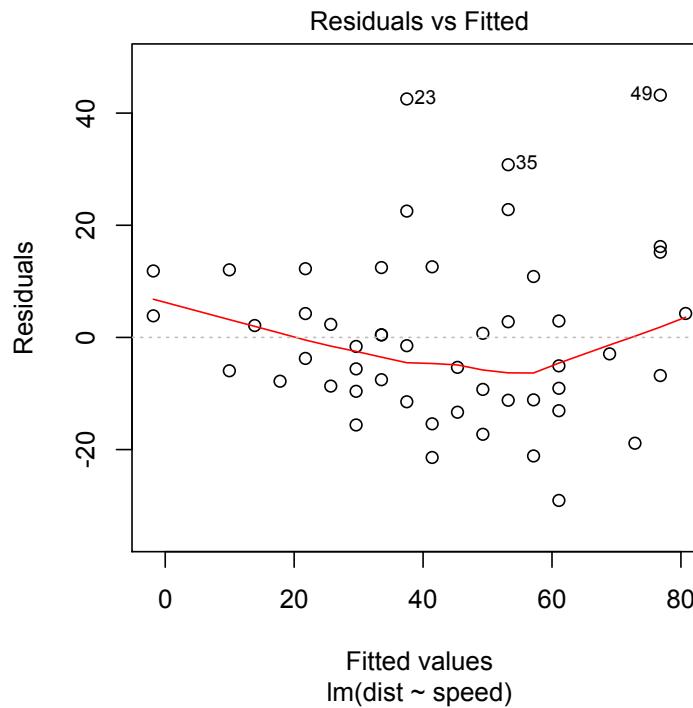
Example: Stopping distance of cars versus speed (mph)



Response: $y = \text{distance}$
Predictor: $x = \text{speed}$

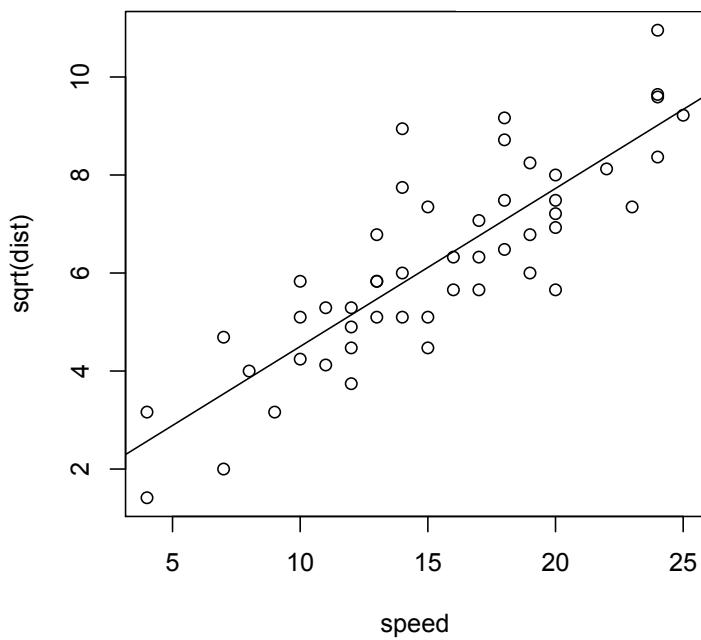
$$y = \alpha + \beta x + \varepsilon$$

$$R^2 = 0.651$$



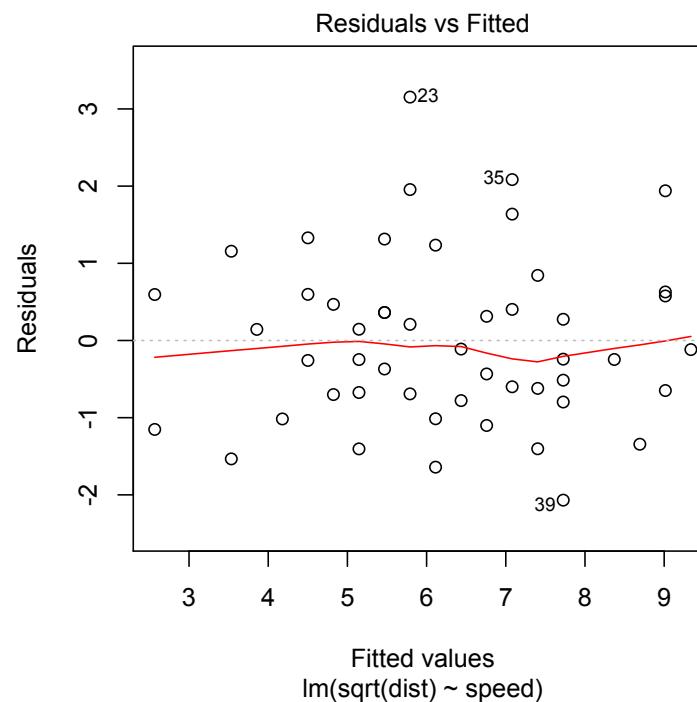
Modelling Non-Linear Relationships

Example: Stopping distance of cars versus speed (mph)



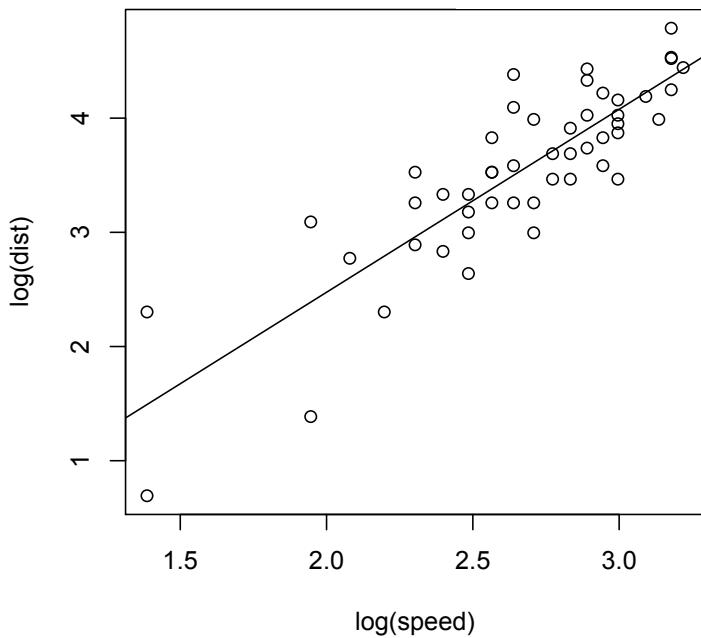
Response: $y = \text{distance}$
Predictor: $x = \text{speed}$

$$y = \alpha + \beta x + \varepsilon \quad R^2 = 0.651$$
$$\sqrt{y} = \alpha + \beta x + \varepsilon \quad R^2 = 0.709$$



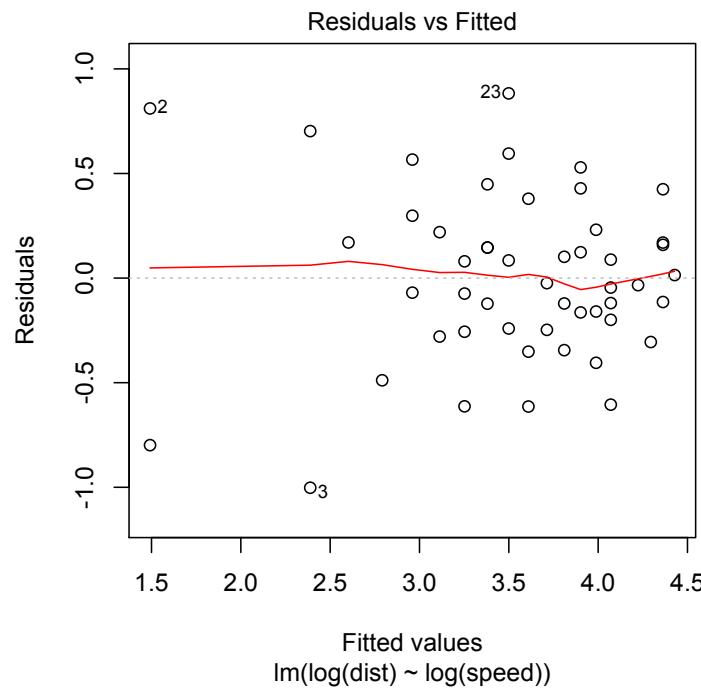
Modelling Non-Linear Relationships

Example: Stopping distance of cars versus speed (mph)



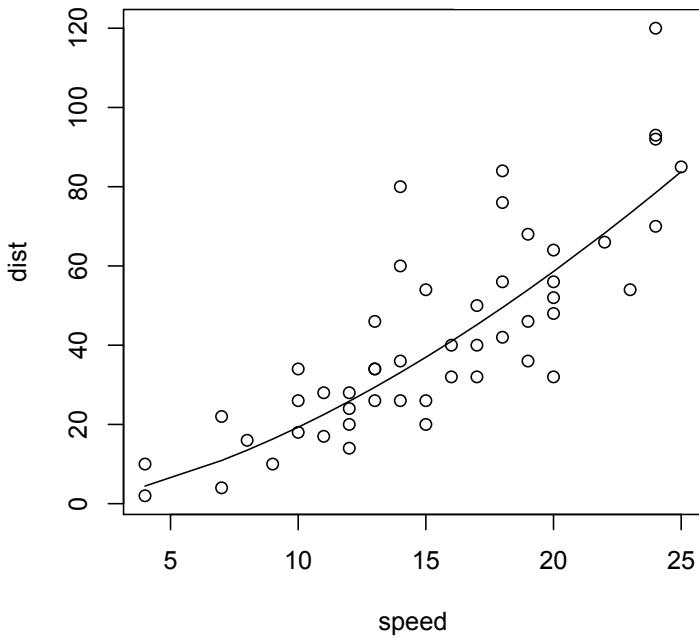
Response: $y = \text{distance}$
Predictor: $x = \text{speed}$

$$y = \alpha + \beta x + \varepsilon \quad R^2 = 0.651$$
$$\sqrt{y} = \alpha + \beta x + \varepsilon \quad R^2 = 0.709$$
$$\log(y) = \alpha + \beta \log(x) + \varepsilon \quad R^2 = 0.733$$



Modelling Non-Linear Relationships

Example: Stopping distance of cars versus speed (mph)



Response: $y = \text{distance}$
Predictor: $x = \text{speed}$

$$y = \alpha + \beta x + \varepsilon \quad R^2 = 0.651$$
$$\sqrt{y} = \alpha + \beta x + \varepsilon \quad R^2 = 0.709$$
$$\log(y) = \alpha + \beta \log(x) + \varepsilon \quad R^2 = 0.733$$

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

Modelling Non-Linear Relationships

Can you use simple regression to fit this model?

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

Non-linear
Multiplicative error model

Modelling Non-Linear Relationships

Can you use simple regression to fit this model?

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

Non-linear
Multiplicative error model

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

Modelling Non-Linear Relationships

Can you use simple regression to fit this model?

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

Non-linear
Multiplicative error model

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

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

Error model is log-Normal.

