

# Linear models in R

John Joseph Valletta

University of Exeter, Penryn Campus, UK

November 2018



Researcher  
Development



# What is a model and why do we need one?

A **model** is a human construct/abstraction that tries to approximate the **data generating process** in some useful manner

# What is a model and why do we need one?

A **model** is a human construct/abstraction that tries to approximate the **data generating process** in some useful manner

## Models are built for

- enhancing our understanding of a complex phenomenon
- executing “what if” scenarios
- predicting/forecasting an outcome
- controlling a system

# What is a model and why do we need one?

A **model** is a human construct/abstraction that tries to approximate the **data generating process** in some useful manner

## Models are built for

- enhancing our understanding of a complex phenomenon
- executing “what if” scenarios
- predicting/forecasting an outcome
- controlling a system

# What is a model and why do we need one?

A **model** is a human construct/abstraction that tries to approximate the **data generating process** in some useful manner

## Models are built for

- enhancing our understanding of a complex phenomenon
- executing “what if” scenarios
- predicting/forecasting an outcome
- controlling a system

# What is a model and why do we need one?

A **model** is a human construct/abstraction that tries to approximate the **data generating process** in some useful manner

## Models are built for

- enhancing our understanding of a complex phenomenon
- executing “what if” scenarios
- predicting/forecasting an outcome
- controlling a system

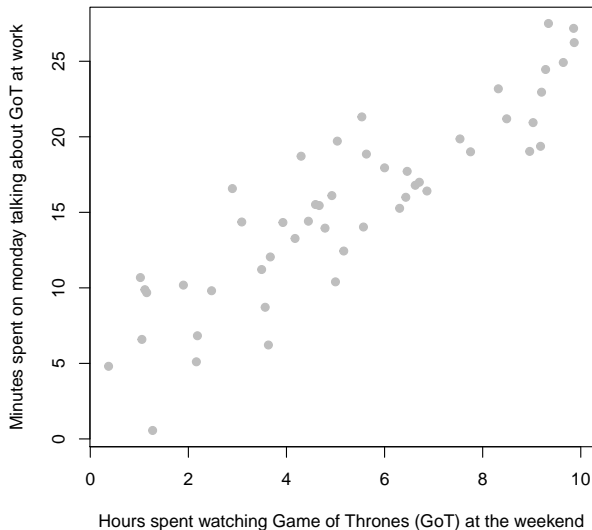
# What is a model and why do we need one?

A **model** is a human construct/abstraction that tries to approximate the **data generating process** in some useful manner

## Models are built for

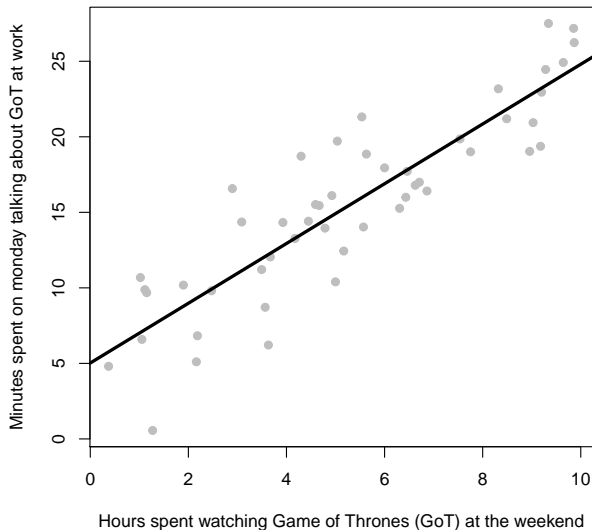
- enhancing our understanding of a complex phenomenon
- executing “what if” scenarios
- predicting/forecasting an outcome
- controlling a system

# Illustrative example

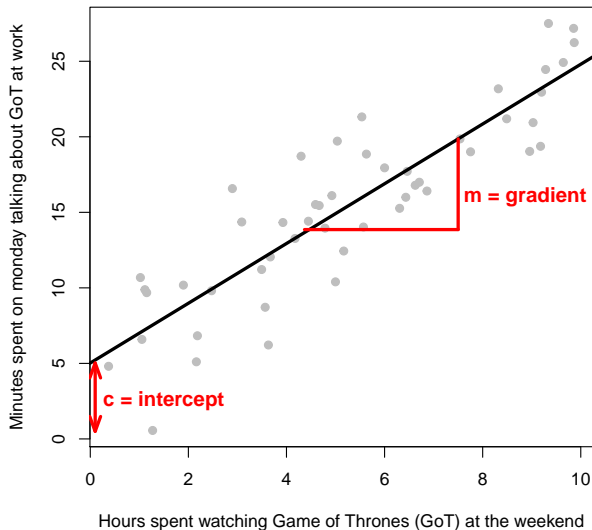




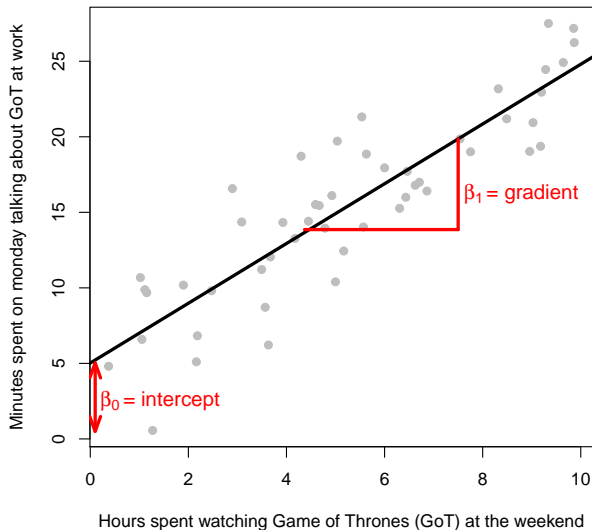
# Illustrative example



# Illustrative example



# Illustrative example



$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

# Formal definition

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

## Observed data

- $y$  (outcome/response): minutes spent talking about GoT
- $x$  (explanatory): hours spent watching Game of Thrones (GoT)

# Formal definition

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

## Observed data

- $y$  (outcome/response): minutes spent talking about GoT
- $x$  (explanatory): hours spent watching Game of Thrones (GoT)

## Parameters to infer

- $\beta_0$ : intercept
- $\beta_1$ : gradient wrt minutes talking about GoT

# Linear models in R

- Use the `lm()` function
- Requires a **formula** object  
`outcome ~ explanatory variable`

```
1 # talk: minutes spent talking about GoT (outcome/response variable)
2 # watch: hours spent watching GoT (explanatory variable)
3
4 fit <- lm(talk ~ watch)
5
6 # If data is in a data frame called "df"
7 fit <- lm(talk ~ watch, df)
```

# Summary of fitted model

```
1 summary(fit)
```

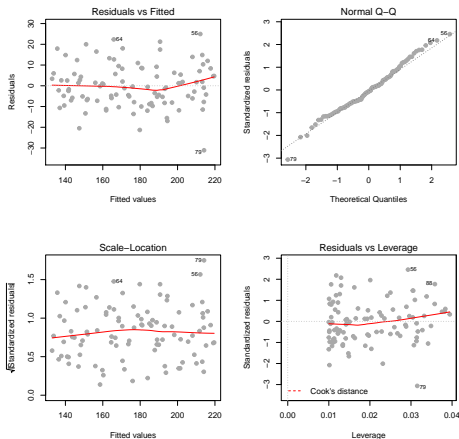
```
##
## Call:
## lm(formula = height ~ weight, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.089  -6.926  -0.689   6.057  24.967
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.35229     7.11668   0.331   0.742
## weight       2.17446     0.08782  24.762 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.31 on 98 degrees of freedom
## Multiple R-squared:  0.8622, Adjusted R-squared:  0.8608
## F-statistic: 613.1 on 1 and 98 DF,  p-value: < 2.2e-16
```



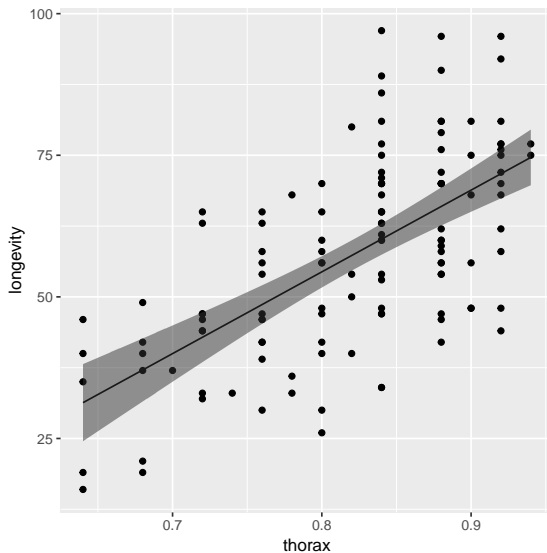
# Model checking

In order to make **robust** inference, we must check the model fit

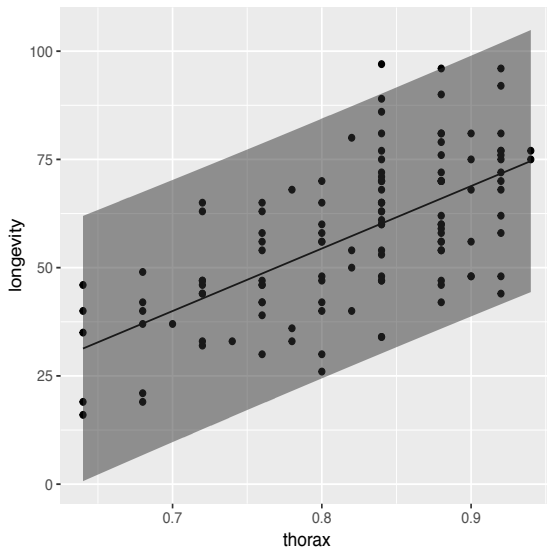
```
1 plot(fit)
```



# Confidence vs prediction intervals

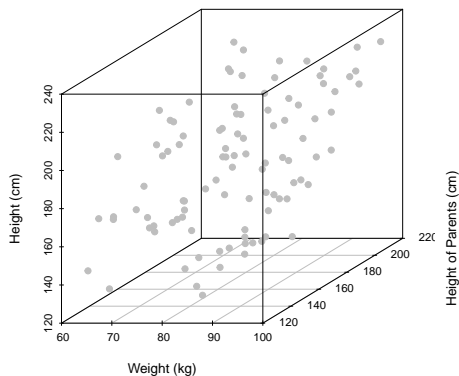


# Confidence vs prediction intervals



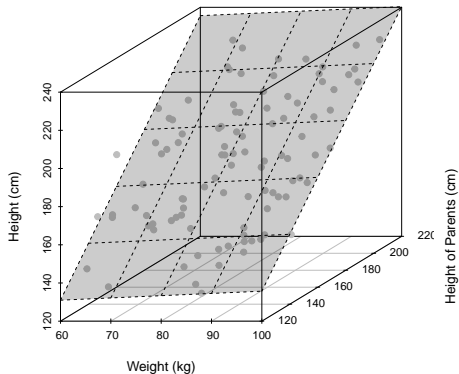
# Multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$



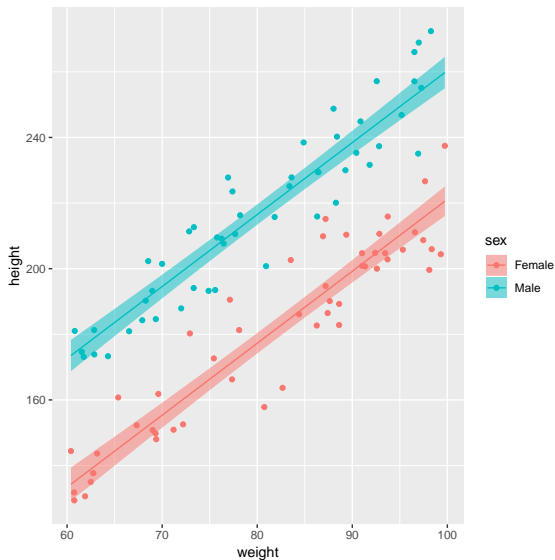
# Multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$



# Categorical variables

# Categorical variables



# Categorical variables

We need **dummy** variables

$$S_i = \begin{cases} 1 & \text{if } i \text{ is male,} \\ 0 & \text{otherwise} \end{cases}$$

Here, female is known as the **baseline/reference level**

The regression is:

$$y_i = \beta_0 + \beta_1 S_i + \beta_2 x_i + \epsilon_i$$

Or in English:

$$\text{height}_i = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{weight}_i + \epsilon_i$$



The mean regression lines for male and female are:

- Female (sex=0)

$$\text{height}_i = \beta_0 + (\beta_1 \times 0) + \beta_2 \text{weight}_i$$

$$\text{height}_i = \beta_0 + \beta_2 \text{weight}_i$$

- Male (sex=1)

$$\text{height}_i = \beta_0 + (\beta_1 \times 1) + \beta_2 \text{weight}_i$$

$$\text{height}_i = (\beta_0 + \beta_1) + \beta_2 \text{weight}_i$$