Linear Models in R

University of Helsinki
Aniol Santo-Angles, November 2025

Linear Models in R

Outline

- Stats
  - Statistical thinking
  - Descriptive Statistics
  - Sampling
  - Normal distribution
  - *Hands-on*
  - Linear models
  - *Hands-on*
- How R works...
  - Data Frames
  - Subset df
  - Plot histogram
  - Fit lm
  - Plot lm outputs
  - Simulate data

Linear Models in R

Outline

- Stats
- **Statistical thinking**
- Descriptive Statistics
- Sampling
- Normal distribution
- *Hands-on*
- Linear models
- *Hands-on*

Linear Models in R

Statistical thinking

- **Statistical thinking** is a way to understanding a complex world by describing it in simple terms while capturing its essential aspects of structure or function (Poldrack, 2023).
- What we can do with statistics?
  - **Describe**: simplify reality in meaningful way (and amenable for our cognitive capabilities).
  - **Decide**: make decisions based on the (limited and noisy) data - uncertainty.
  - **Predict**: project previous experiences to new situations.

Linear Models in R

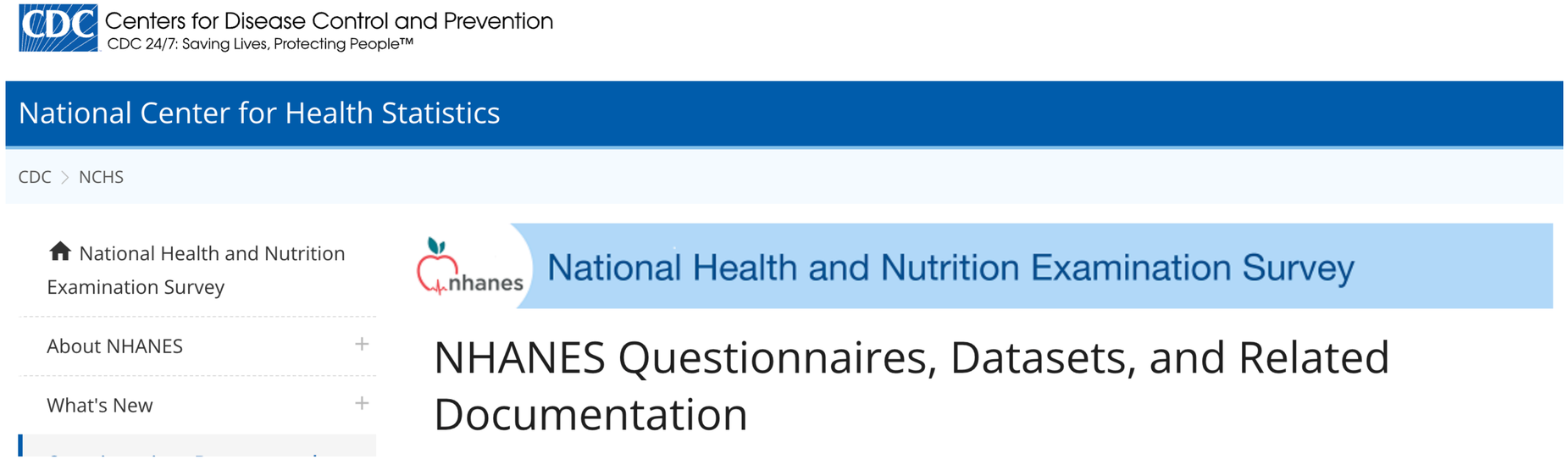
Outline

- Stats
- Statistical thinking
- **Descriptive Statistics**
- Sampling
- Normal distribution
- *Hands-on*
- Linear models
- *Hands-on*

Linear Models in R

Descriptive Statistics

- We can better understand the world by **throwing away information**, this is what descriptive statistics does.



https://wwwn.cdc.gov/nchs/nhanes/

Linear Models in R

Descriptive Statistics

- We can better understand the world by **throwing away information**, this is what descriptive statistics does.

Age	Gender	Height_cm	Weight_kg
2	Female	88.6	13.7
2	Male	94.2	13.9
66	Female	158.3	79.5
18	Male	175.7	66.3
13	Male	158.4	45.4
66	Female	150.2	53.5
75	Female	151.1	88.8
0	Female	*NA*	10.2
56	Male	170.6	62.1
18	Male	172.8	58.9
67	Male	178.6	74.9
54	Female	147.8	87.1
71	Male	170.6	65.6
61	Male	159.2	77.7
22	Male	174.1	74.4
45	Male	157.3	54.4
13	Female	160.5	67.0
1	Female	*NA*	10.6
60	Female	154.0	85.1
60	Female	154.6	56.8
64	Male	170.1	64.9
5	Female	111.4	18.2
9	Female	145.8	34.3
67	Female	154.5	74.3
70	Male	161.9	62.7
53	Male	188.1	118.1
42	Male	173.2	82.8
57	Male	167.6	80.2
20	Male	179.9	89.2
72	Male	162.4	56.1
11	Female	142.6	40.8
52	Male	147.7	76.5
11	Male	143.4	36.9
8	Male	137.6	35.0
26	Male	170.2	97.7
5	Female	116.3	22.5
72	Male	183.4	102.9
9	Female	135.8	31.9
72	Male	178.3	107.9
61	Male	187.8	79.3
10	Female	154.7	50.8

Central Tendency

$$\text{Mean:} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Median:} \quad \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & n \text{ even} \end{cases}$$

$$\text{Mode:} \quad \text{mode}(x) = \arg\max_{x} f(x)$$

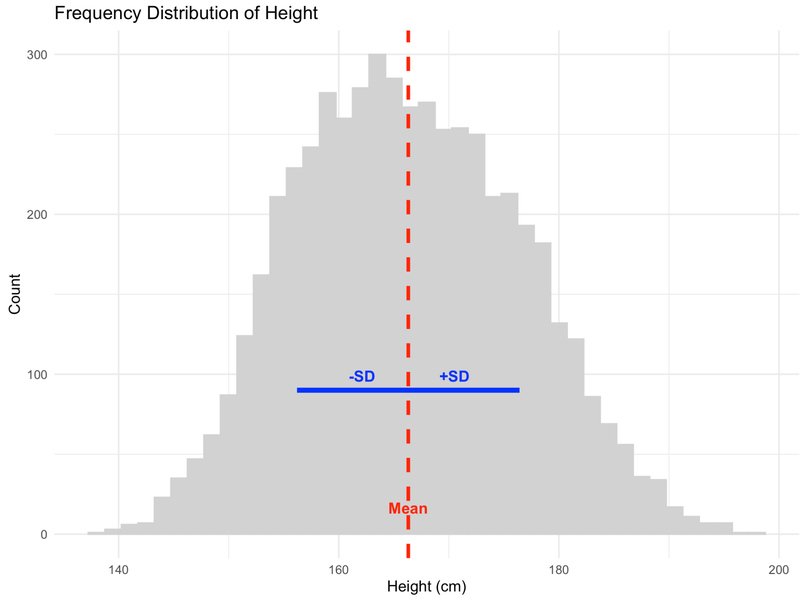
NHANES dataset (n= 8000)

Linear Models in R

Descriptive Statistics

- We can better understand the world by **throwing away information**, this is what descriptive statistics does.

Age	Gender	Height_cm	Weight_kg
2	Female	88.6	13.7
2	Male	94.2	13.9
66	Female	158.3	79.5
18	Male	175.7	66.3
13	Male	158.4	45.4
66	Female	150.2	53.5
75	Female	151.1	88.8
0	Female	*NA*	10.2
56	Male	170.6	62.1
18	Male	172.8	58.9
67	Male	178.6	74.9
54	Female	147.8	87.1
71	Male	170.6	65.6
61	Male	159.2	77.7
22	Male	174.1	74.4
45	Male	157.3	54.4
13	Female	160.5	67.0
1	Female	*NA*	10.6
60	Female	154.0	85.1
60	Female	154.6	56.8
64	Male	170.1	64.9
5	Female	111.4	18.2
9	Female	145.8	34.3
67	Female	154.5	74.3
70	Male	161.9	62.7
53	Male	188.1	118.1
42	Male	173.2	82.8
57	Male	167.6	80.2
20	Male	179.9	89.2
72	Male	162.4	56.1
11	Female	142.6	40.8
52	Male	147.7	76.5
11	Male	143.4	36.9
8	Male	137.6	35.0
26	Male	170.2	97.7
5	Female	116.3	22.5
72	Male	183.4	102.9
9	Female	135.8	31.9
72	Male	178.3	107.9
61	Male	187.8	79.3
10	Female	154.7	50.8



NHANES dataset (n= 8000)

Central Tendency

$$\text{Mean:} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Median:} \quad \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & n \text{ even} \end{cases}$$

$$\text{Mode:} \quad \text{mode}(x) = \arg\max_x f(x)$$

Dispersion - Spread

$$\text{Variance: } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Standard Deviation: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

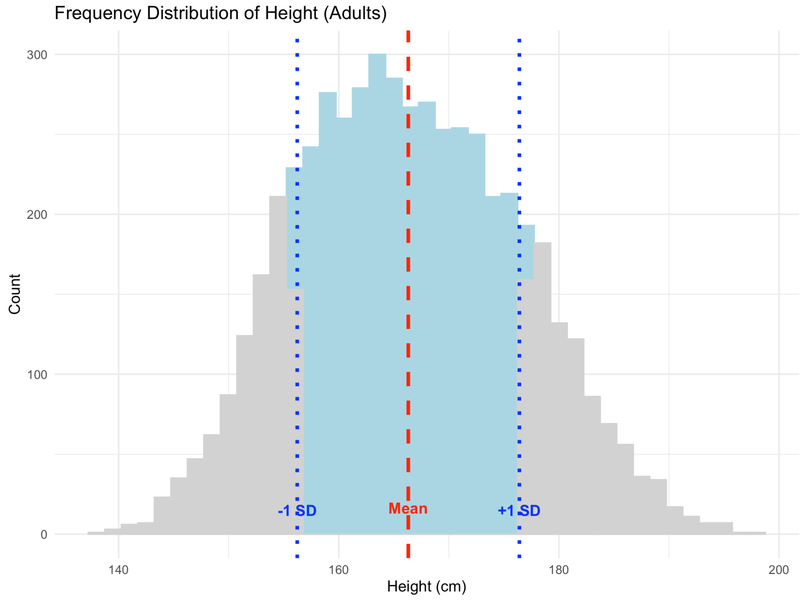
Linear Models in R

Descriptive Statistics

- We can better understand the world by **throwing away information**, this is what descriptive statistics does.

Age	Gender	Height_cm	Weight_kg
2	Female	88.6	13.7
2	Male	94.2	13.9
66	Female	158.3	79.5
18	Male	175.7	66.3
13	Male	158.4	45.4
66	Female	150.2	53.5
75	Female	151.1	88.8
0	Female	*NA*	10.2
56	Male	170.6	62.1
18	Male	172.8	58.9
67	Male	178.6	74.9
54	Female	147.8	87.1
71	Male	170.6	65.6
61	Male	159.2	77.7
22	Male	174.1	74.4
45	Male	157.3	54.4
13	Female	160.5	67.0
1	Female	*NA*	10.6
60	Female	154.0	85.1
60	Female	154.6	56.8
64	Male	170.1	64.9
5	Female	111.4	18.2
9	Female	145.8	34.3
67	Female	154.5	74.3
70	Male	161.9	62.7
53	Male	188.1	118.1
42	Male	173.2	82.8
57	Male	167.6	80.2
20	Male	179.9	89.2
72	Male	162.4	56.1
11	Female	142.6	40.8
52	Male	147.7	76.5
11	Male	143.4	36.9
8	Male	137.6	35.0
26	Male	170.2	97.7
5	Female	116.3	22.5
72	Male	183.4	102.9
9	Female	135.8	31.9
72	Male	178.3	107.9
61	Male	187.8	79.3
10	Female	154.7	50.8

NHANES dataset (n= 8000)



68% observations (normal distribution)

Central Tendency

$$\text{Mean:} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Median:} \quad \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & n \text{ even} \end{cases}$$

$$\text{Mode:} \quad \text{mode}(x) = \arg\max_x f(x)$$

Dispersion - Spread

$$\text{Variance: } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Standard Deviation: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

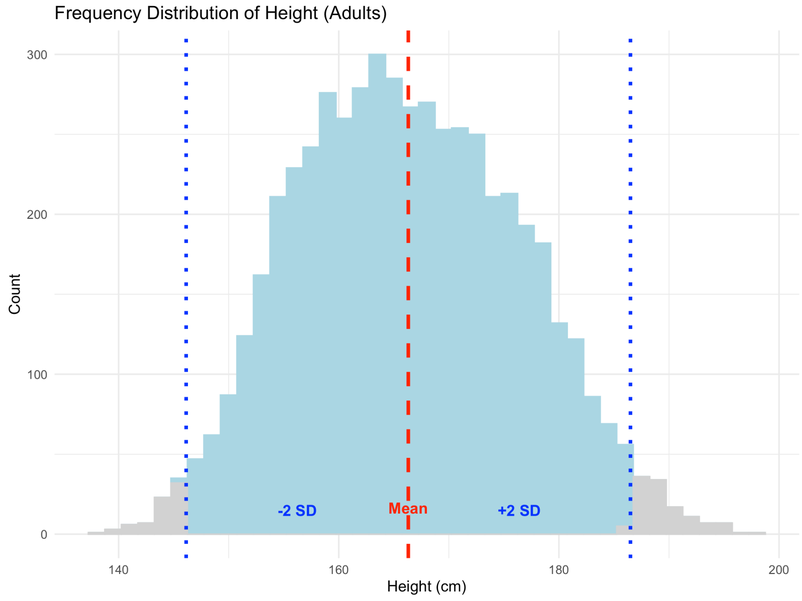
Linear Models in R

Descriptive Statistics

- We can better understand the world by **throwing away information**, this is what descriptive statistics does.

Age	Gender	Height_cm	Weight_kg
2	Female	88.6	13.7
2	Male	94.2	13.9
66	Female	158.3	79.5
18	Male	175.7	66.3
13	Male	158.4	45.4
66	Female	150.2	53.5
75	Female	151.1	88.8
0	Female	*NA*	10.2
56	Male	170.6	62.1
18	Male	172.8	58.9
67	Male	178.6	74.9
54	Female	147.8	87.1
71	Male	170.6	65.6
61	Male	159.2	77.7
22	Male	174.1	74.4
45	Male	157.3	54.4
13	Female	160.5	67.0
1	Female	*NA*	10.6
60	Female	154.0	85.1
60	Female	154.6	56.8
64	Male	170.1	64.9
5	Female	111.4	18.2
9	Female	145.8	34.3
67	Female	154.5	74.3
70	Male	161.9	62.7
53	Male	188.1	118.1
42	Male	173.2	82.8
57	Male	167.6	80.2
20	Male	179.9	89.2
72	Male	162.4	56.1
11	Female	142.6	40.8
52	Male	147.7	76.5
11	Male	143.4	36.9
8	Male	137.6	35.0
26	Male	170.2	97.7
5	Female	116.3	22.5
72	Male	183.4	102.9
9	Female	135.8	31.9
72	Male	178.3	107.9
61	Male	187.8	79.3
10	Female	154.7	50.8

NHANES dataset (n= 8000)



95% observations (normal distribution)

Central Tendency

$$\text{Mean:} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Median:} \quad \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & n \text{ even} \end{cases}$$

$$\text{Mode:} \quad \text{mode}(x) = \arg\max_x f(x)$$

Dispersion - Spread

$$\text{Variance: } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Standard Deviation: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Linear Models in R

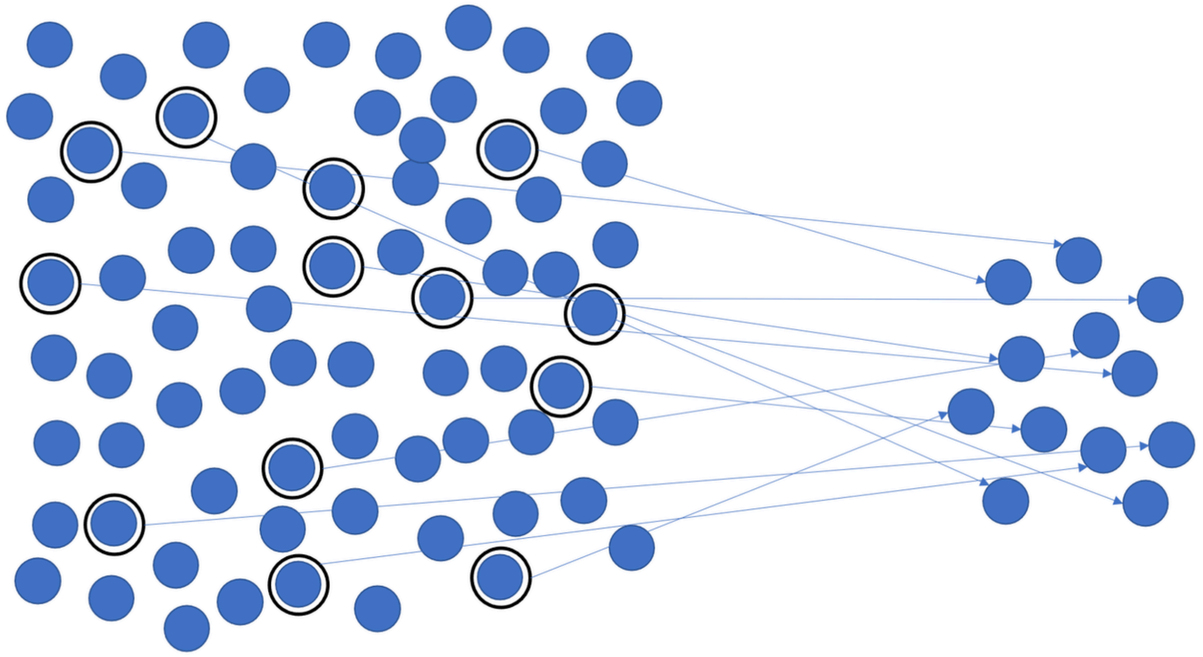
Outline

- Stats
- Statistical thinking
- Descriptive Statistics
- **Sampling**
- Normal distribution
- *Hands-on*
- Linear models
- *Hands-on*

Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.



Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

	Population	*Sample*
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

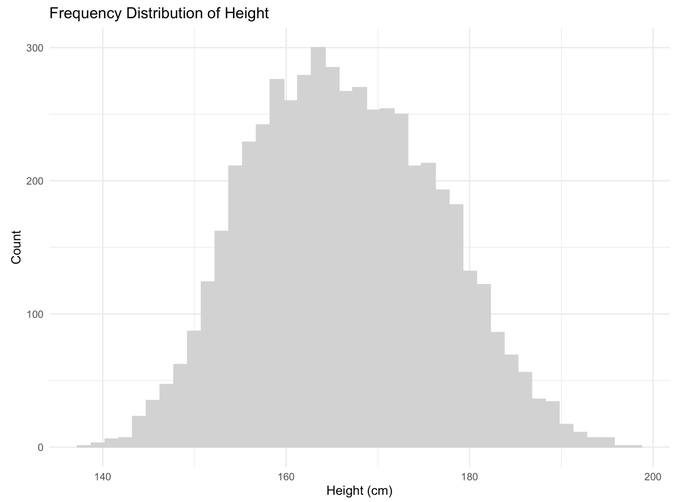
?

Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

Population is Height NHANES dataset (n=5311)



500 samples of 50 observations

SampleID	PopulationMean	SampleMean
1	166.3163	165.584
2	166.3163	165.634
3	166.3163	167.120
4	166.3163	166.394
5	166.3163	165.542

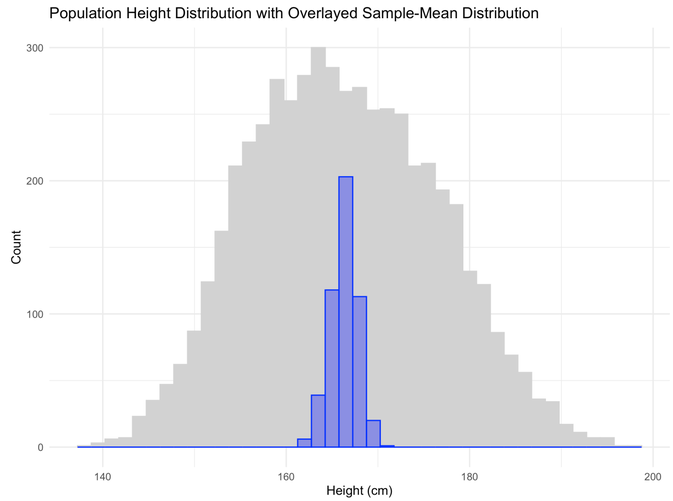
***Sampling error**: difference between population and sample estimate (e.g. mean)*

Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

Population is Height NHANES dataset (n=5311)



500 samples of 50 observations

SampleID	PopulationMean	SampleMean
1	166.3163	165.584
2	166.3163	165.634
3	166.3163	167.120
4	166.3163	166.394
5	166.3163	165.542

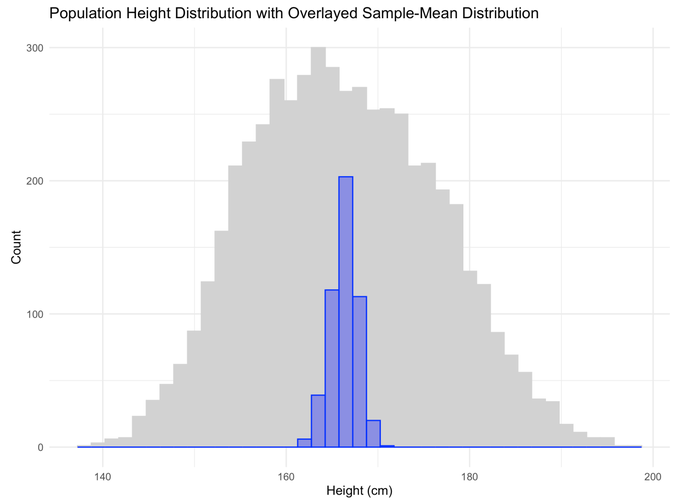
Sampling distribution of the sample mean
Distribution of sample means

Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

Population is Height NHANES dataset (n=5311)



Sampling distribution of the sample mean
Distribution of sample means

***Standard Error of the Mean (SEM)**:*

How much the sample mean changes

How uncertain our estimate of the mean is

$$S.E.M. = \frac{\hat{\sigma}}{\sqrt{n}}$$

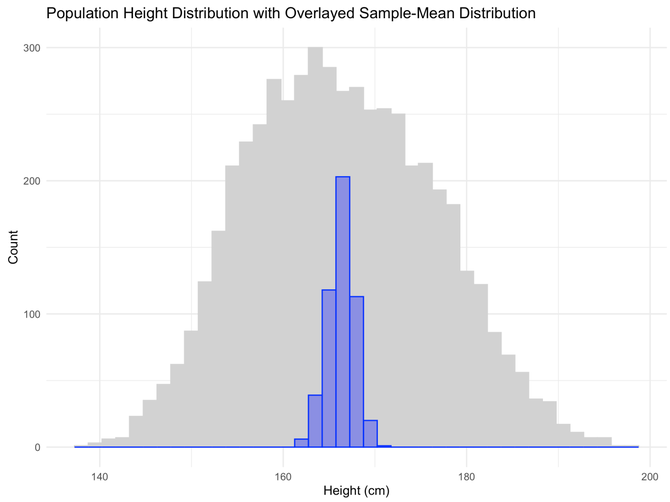
SEM is the standard deviation of the sampling distribution of the mean

Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

Population is Height NHANES dataset (n=5311)



Sampling distribution of the sample mean
Distribution of sample means

***Standard Error of the Mean (SEM)**:*

How much the sample mean changes

How uncertain our estimate of the mean is

$$S.E.M. = \frac{\hat{\sigma}}{\sqrt{n}}$$

SEM is the standard deviation of the sampling distribution of the mean

***Confidence Interval**:*

range of plausible values for the true population mean

$$C.I. = \bar{x} \pm t_{critic} * S.E.M$$

$$C.I. = \text{point estimate} \pm \text{critical value} * \text{standard error}$$

Interpretation:

WRONG: 95% chance that the population mean lies within the interval — the population mean is fixed!

CORRECT: 95% of the intervals computed from those samples would contain the true population mean.

Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

	Population	*Sample*
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Degrees of freedom

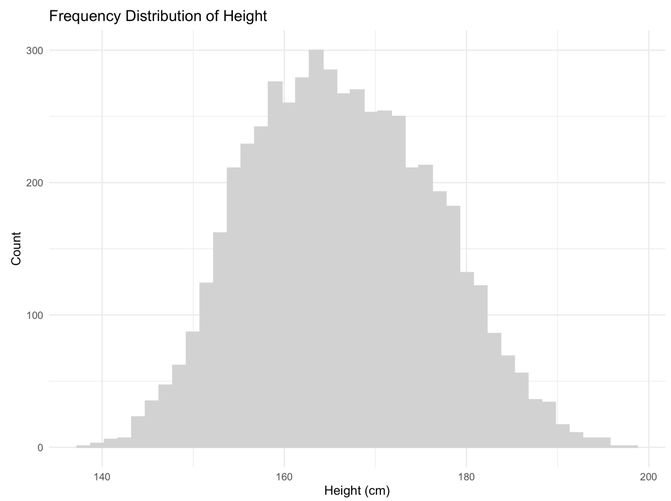
best estimate of the population variance (σ^2) is the sample variance (s^2), and this is thanks to the d.f. correction.

Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

Population is Height NHANES dataset (n=5311)



500 samples of 30 observations

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

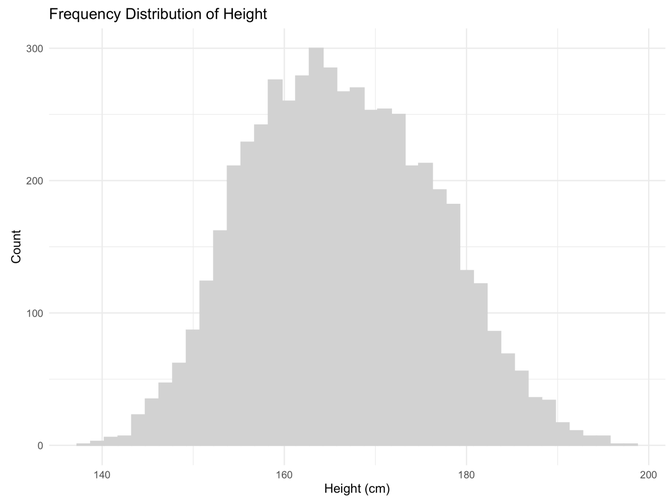
SampleID	TrueVar	SampleVar	PopVar	Closest
1	101.805	78.02351	75.42272	SampleVar
2	101.805	90.85426	87.82579	SampleVar
3	101.805	92.45220	89.37046	SampleVar
4	101.805	119.39375	115.41396	PopVar
5	101.805	94.86944	91.70712	SampleVar
6	101.805	88.90392	85.94046	SampleVar
7	101.805	106.52920	102.97822	PopVar
8	101.805	77.65430	75.06582	SampleVar
9	101.805	120.53237	116.51462	PopVar
10	101.805	60.91771	58.88712	SampleVar

Linear Models in R

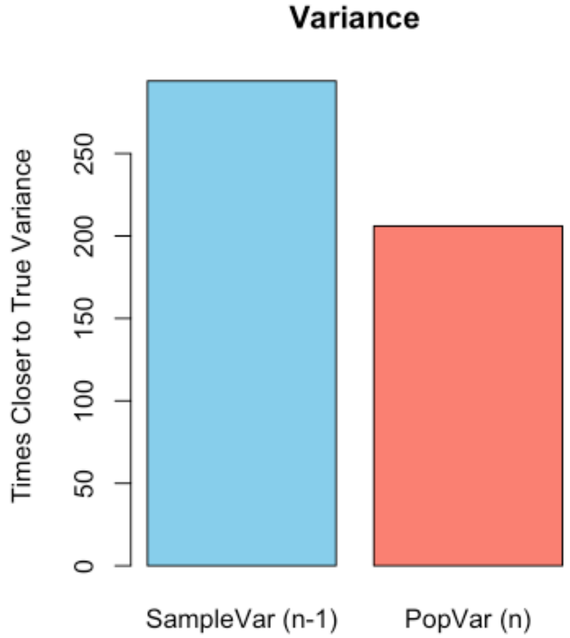
Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

Population is Height NHANES dataset (n=5311)



500 samples of 30 observations

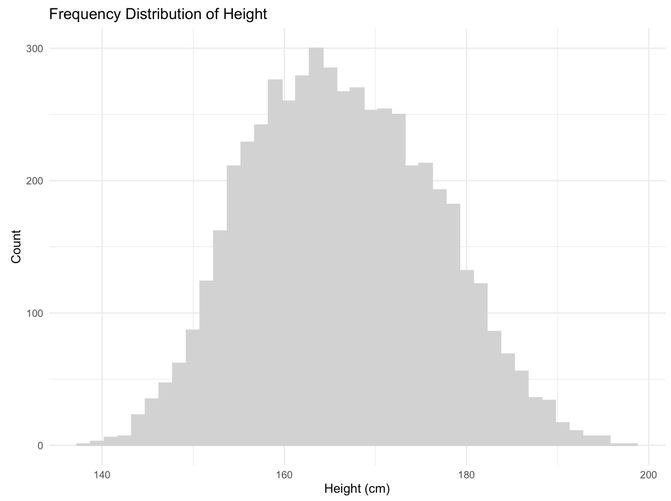


Linear Models in R

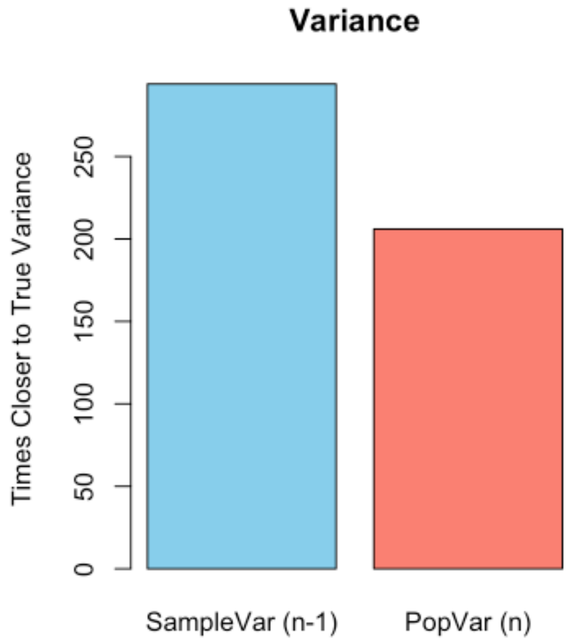
Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

Population is Height NHANES dataset (n=5311)



500 samples of 30 observations



Degrees of freedom

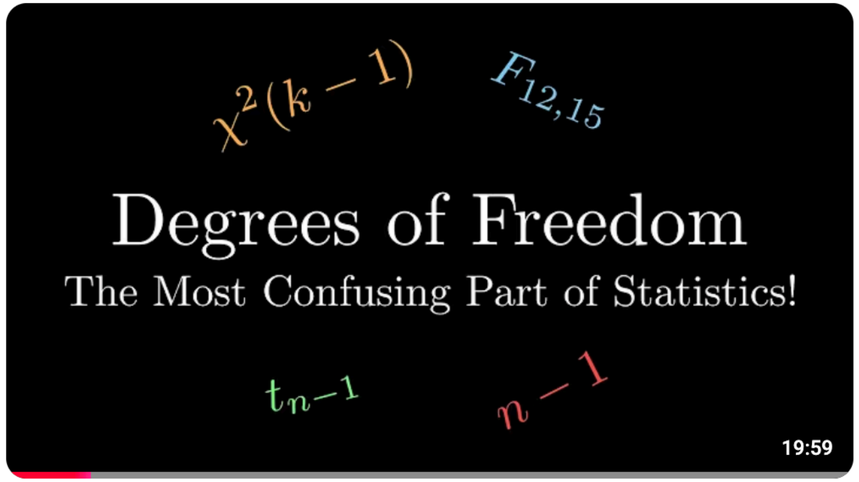
best estimate of the population variance (σ^2) is the sample variance (s^2), and this is thanks to the d.f. correction.

Linear Models in R

Sample and Population

- Sampling from a population to make inferences about the population as a whole, not just the sample.

Degrees of freedom: WHY?



Degrees of Freedom, Actually Explained - The Geometry of Statistics | Ch. 1 (#SoME4)

161K views • 4 months ago

Sam Levey

The most confusing concept in statistics must be **degrees of freedom**. Students everywhere leave their introductory stats courses ...

Overview

14 chapters Introduction | Basic Definition | The Tweet That Started It All | The Geometry of Statistics |...

Linear Models in R

Outline

- Stats
- Statistical thinking
- Descriptive Statistics
- Sampling
- **Normal distribution**
- *Hands-on*
- Linear models
- *Hands-on*

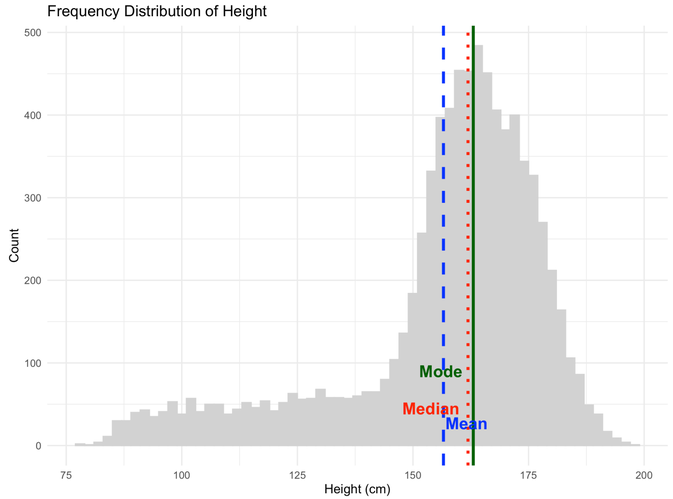
Linear Models in R

Normal distribution

- Normal distribution is fully described by 1st and 2nd order statistics (mean and std).
- Other distributions, 3rd order statistics are needed (skewed, kurtosis,...)

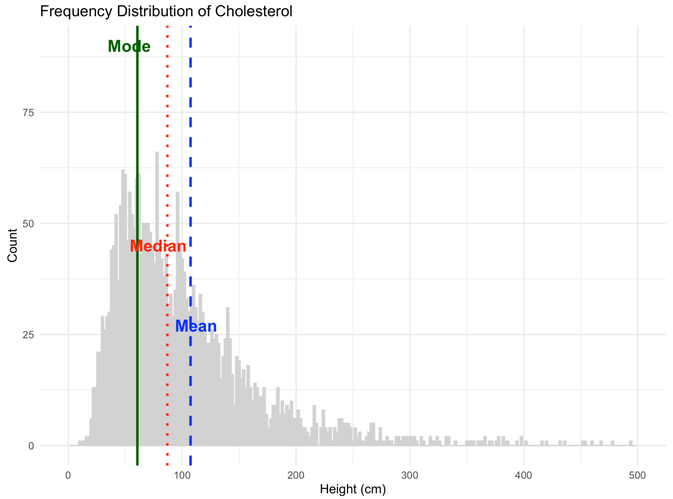
Height across adult population

Normal distribution



Height across all population

Left-skewed Distribution



Cholesterol across all population

Right-skewed Distribution

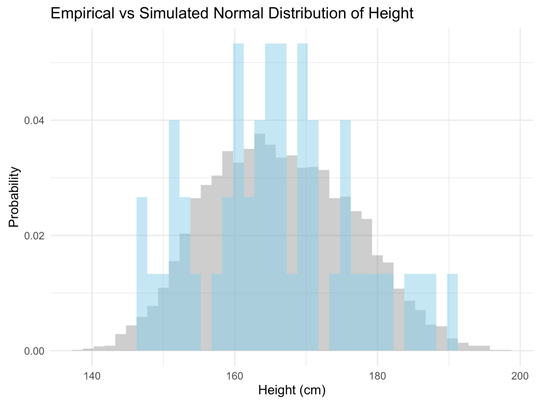
Long-tailed (heavy-tailed) distribution

Linear Models in R

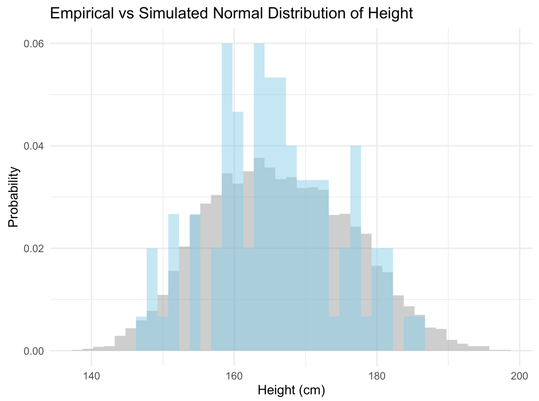
Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

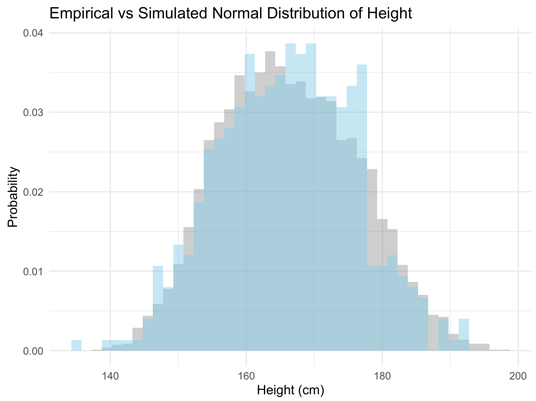
50 samples



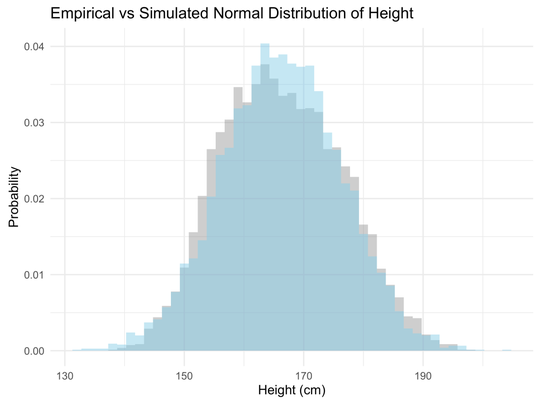
100 samples



500 samples



5000 samples

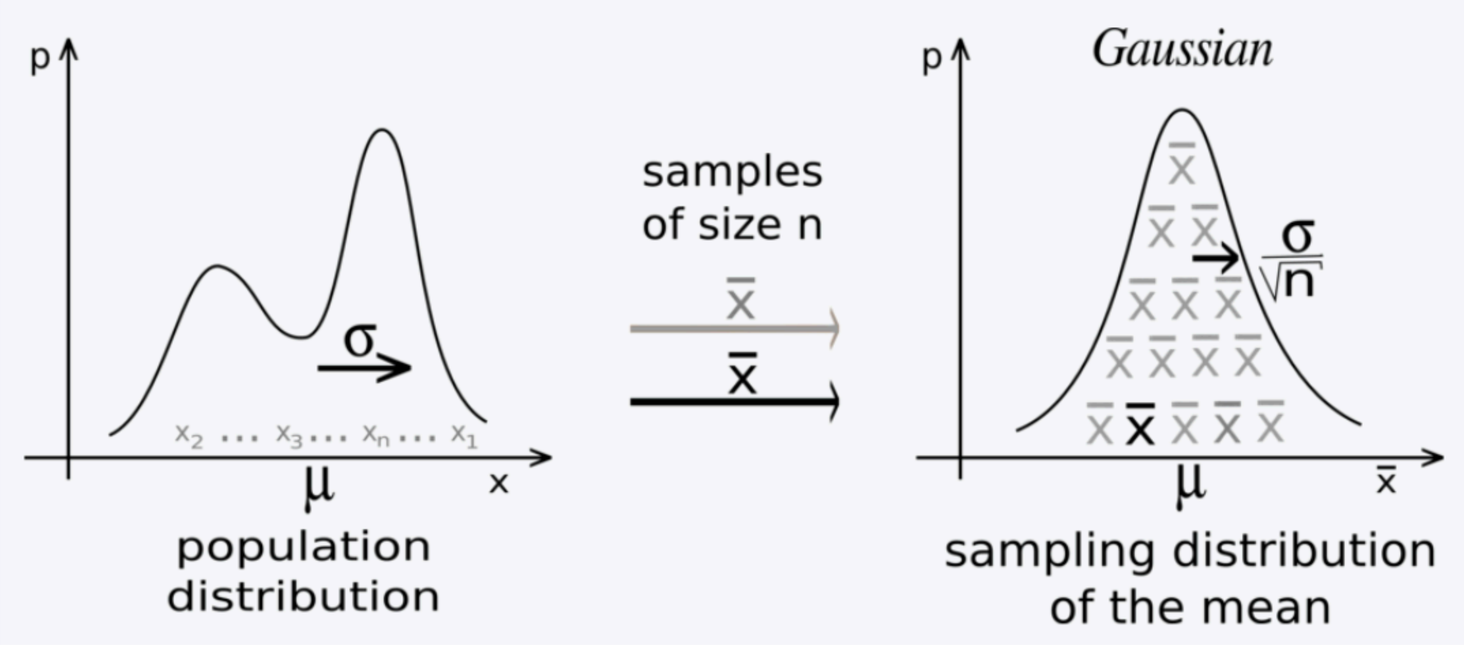


Empirical
Simulated

Linear Models in R

Normal distribution

- Why does the normal distribution occur so frequently in nature and data? **Central Limit Theorem.**
- The sampling distribution of the mean approaches normal distribution, even if the original variables are not normal



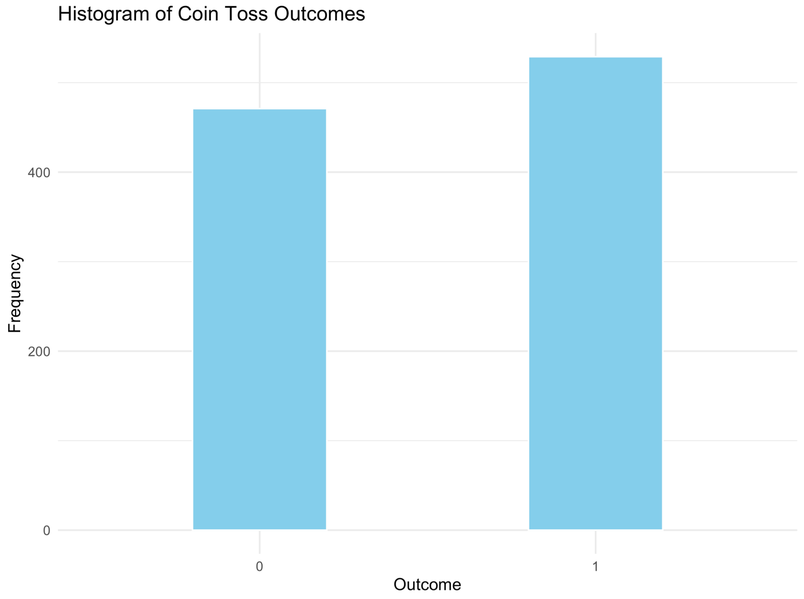
Wikipedia

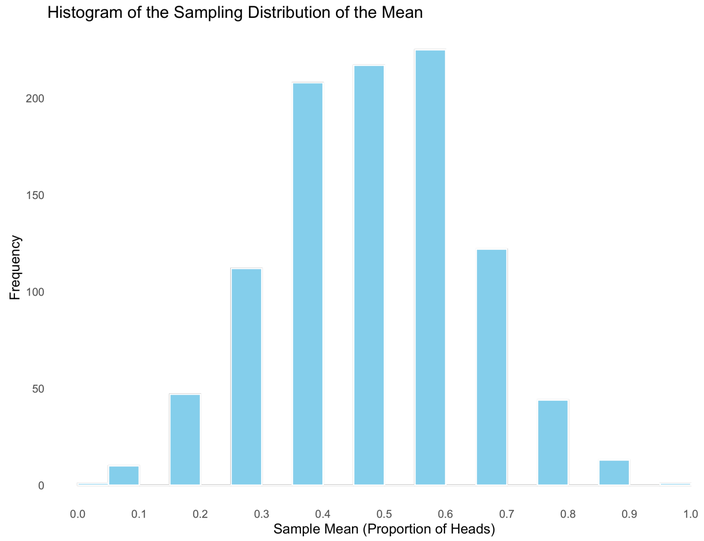
Linear Models in R

Normal distribution

- Why does the normal distribution occur so frequently in nature and data? **Central Limit Theorem.**
- The sampling distribution of the mean approaches normal distribution, even if the original variables are not normal

Simulate 1000 tosses of a fair coin (P(heads) = 0.5), and plot histogram.





Simulate 1000 tosses of a fair coin (P(heads) = 0.5), compute the mean number of heads in groups of 10 tosses, and plot the resulting sampling distribution of the mean.

Hands-on

github.com/asantoangles/linear_models_in_R

script_01_linear_models.R

Linear Models in R

Outline

- Stats
  - Statistical thinking
  - Descriptive Statistics
  - Sampling
  - Normal distribution
  - *Hands-on*
  - **Linear models**
  - *Hands-on*

Linear Models in R

Statistical model

- “All models are wrong, but some are useful” (George Box, 1976... but heard from someone else...).
- “The best model of a cat is another cat, preferentially the same cat” (Norbert Wiener, 1945).
- Models are **useful simplifications** of the world (or a part of it).
- Statistical models are **theories about how the data we observe is generated**, and **statistical modeling** is the process of finding the most accurate theory.

Linear Models in R

Statistical model

$$data = model + error$$

$$\widehat{data}_i = model_i$$

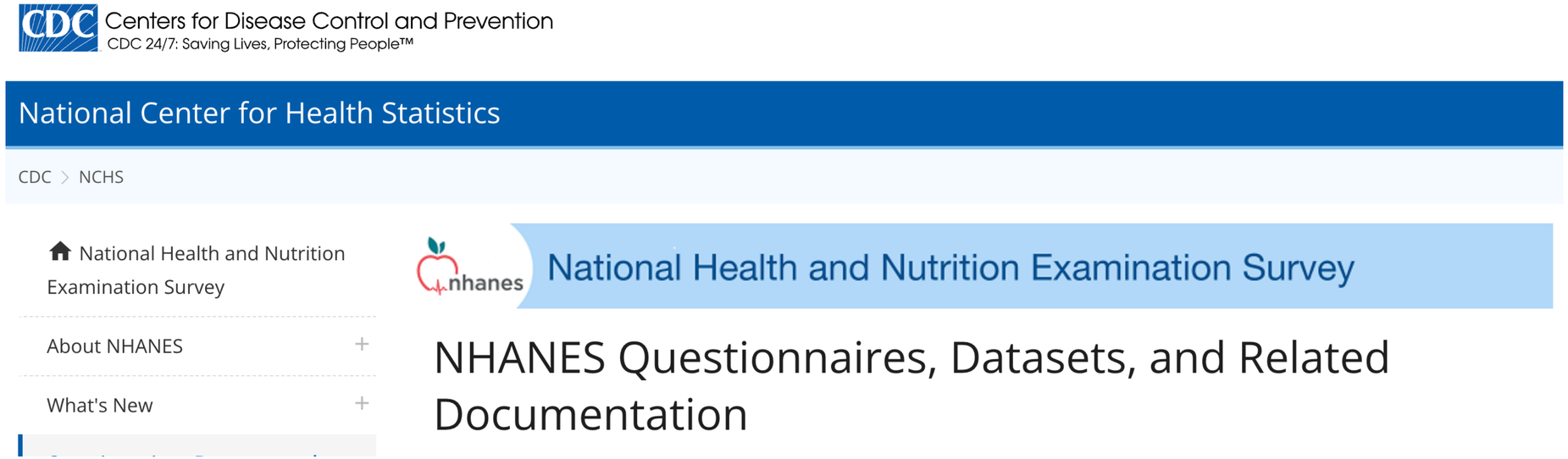
$$error_i = data_i - \widehat{data}_i$$

- Once model is fitted (see later), the model outputs estimates (data_hat)
- Error (or residuals) is variance not explained.
- Goal: test whether model explains data properly, that is, minimize error

Linear Models in R

Statistical model

- How Can We Explain Why People Have Different Heights?



https://wwwn.cdc.gov/nchs/nhanes/

Linear Models in R

Statistical model

- Model 0 - model height as a function of height mean

$$height = \overline{height} + error$$

$$y_i = \bar{y} + \epsilon$$

$$y_i = \hat{\beta} + \epsilon$$

Response Variable

(Vector)

Predictor Variable

(Scalar)

$$error_i = y_i - \hat{y}_i$$

General form

$$error_i = y_i - \hat{\beta}$$

$$error_i = y_i - \bar{y}$$

Specific of this example

Linear Models in R

Statistical model

- Model 1 - model height as a function of age

$$height = slope * age + error$$

$$y_i = \hat{\beta}_1 * x_i + \epsilon$$

Response Variable

(Vector)

Regression coefficient

(Scalar)

Predictor Variable

(Vector)

$$error_i = y_i - \hat{y}_i$$

General form

$$error_i = y_i - \hat{\beta}_1 * x_i$$

Specific of this model

Fitting the model means finding the value of B_1 that makes the model explain most of the variance (minimize error!)

Linear Models in R

Statistical model

- Model 1 - model height as a function of age

$$height = slope * age + error$$

$$y_i = \hat{\beta}_1 * x_i + \epsilon$$

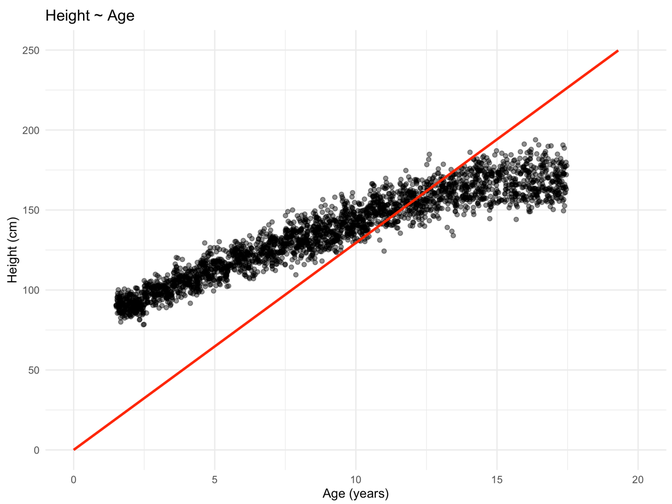
Linear Models in R

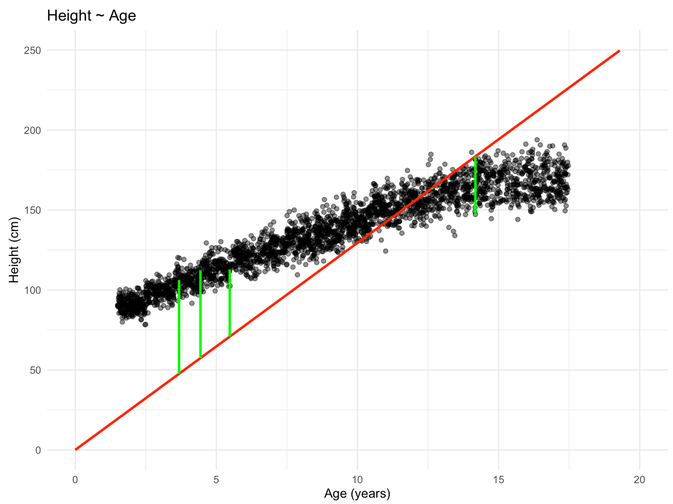
Statistical model

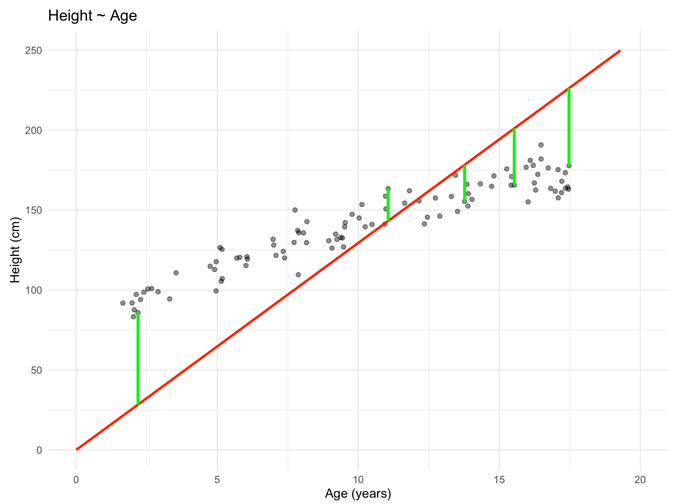
$$height = slope * age + error$$

$$y_i = \hat{\beta}_1 * x_i + \epsilon$$

- Model 1 - model height as a function of age







Model explains poorly the data... why?

At age zero, the model predicts height is zero—so the regression line starts at the origin (x=0, y=0)!

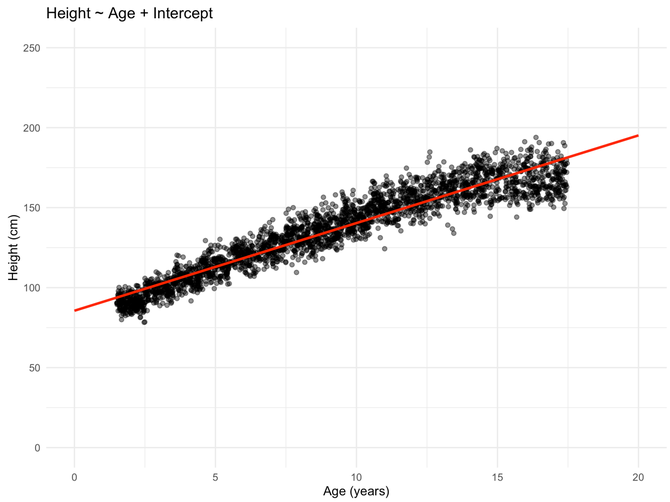
Linear Models in R

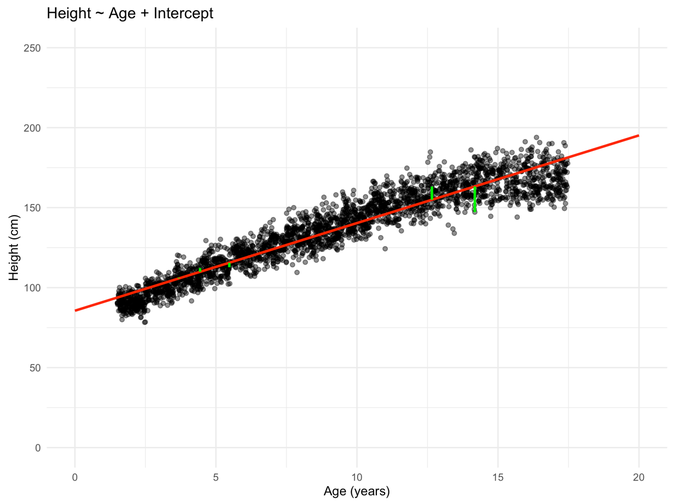
Statistical model

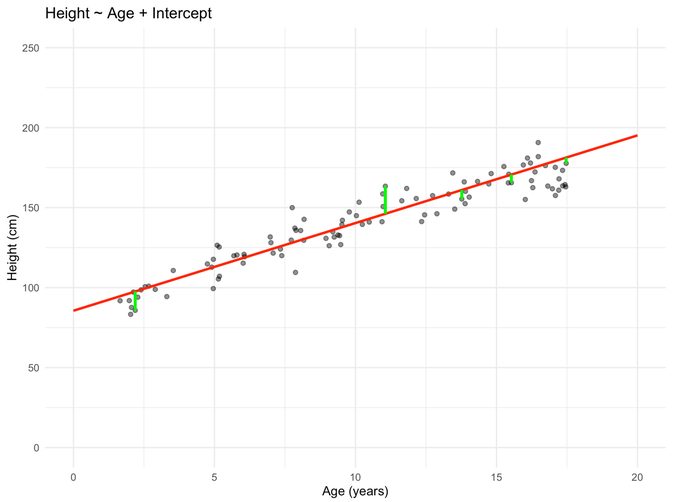
$$height = intercept + slope * age + error$$

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i + \epsilon$$

- Model 2 - model height as a function of age, with intercept



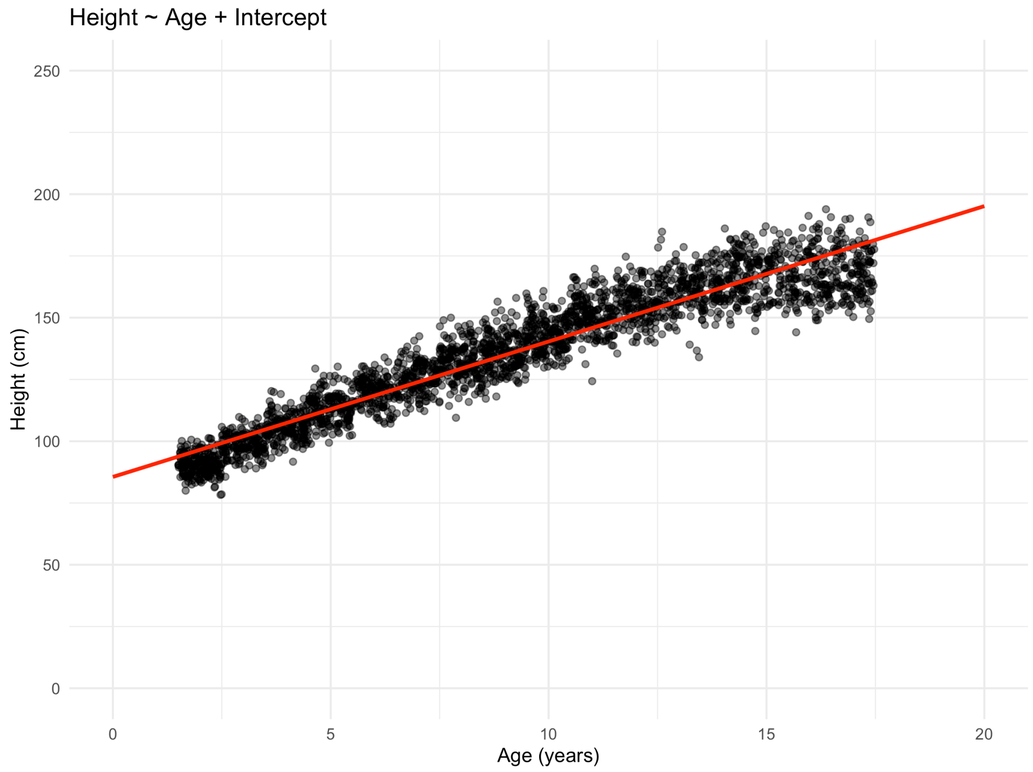




Linear Models in R

Statistical model

- Model 2 - model height as a function of age, with intercept

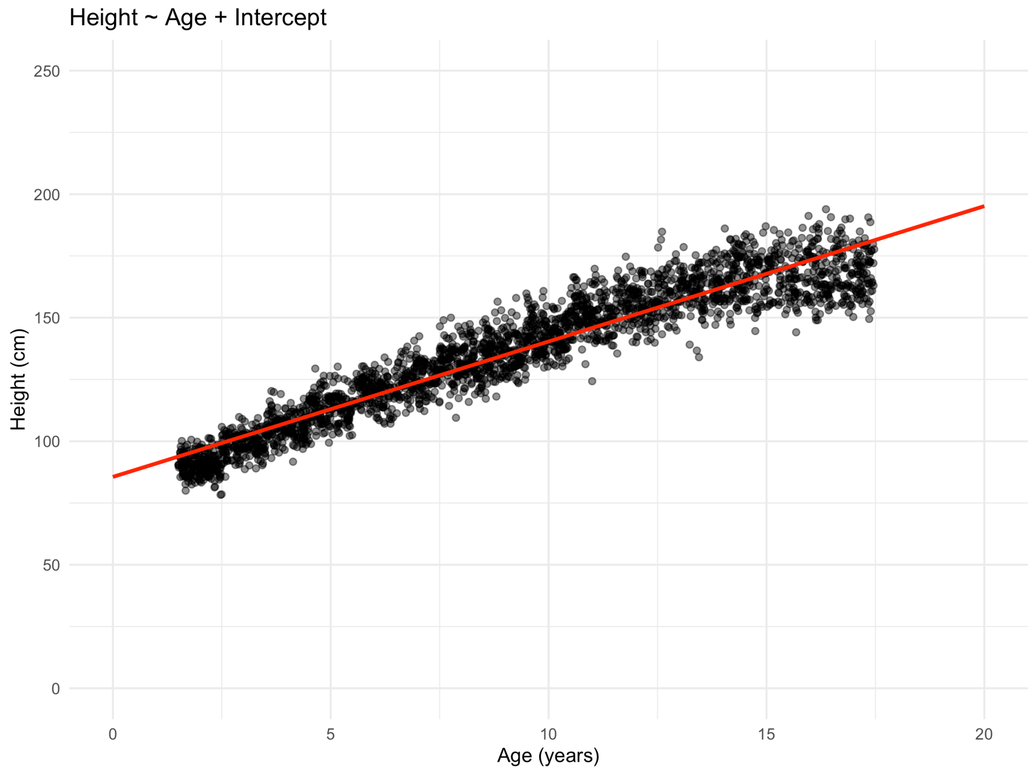



```
Call:
lm(formula = Height_cm ~ Age, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1814  -5.4882   0.0058   5.5853  30.1664

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56411    0.37412   228.7   <2e-16 ***
Age          5.48009    0.03645   150.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.456 on 2570 degrees of freedom
Multiple R-squared:  0.8979,    Adjusted R-squared:  0.8979
F-statistic: 2.26e+04 on 1 and 2570 DF,  p-value: < 2.2e-16
```

Linear Models in R

Statistical model

- Model 2 - model height as a function of age, with intercept

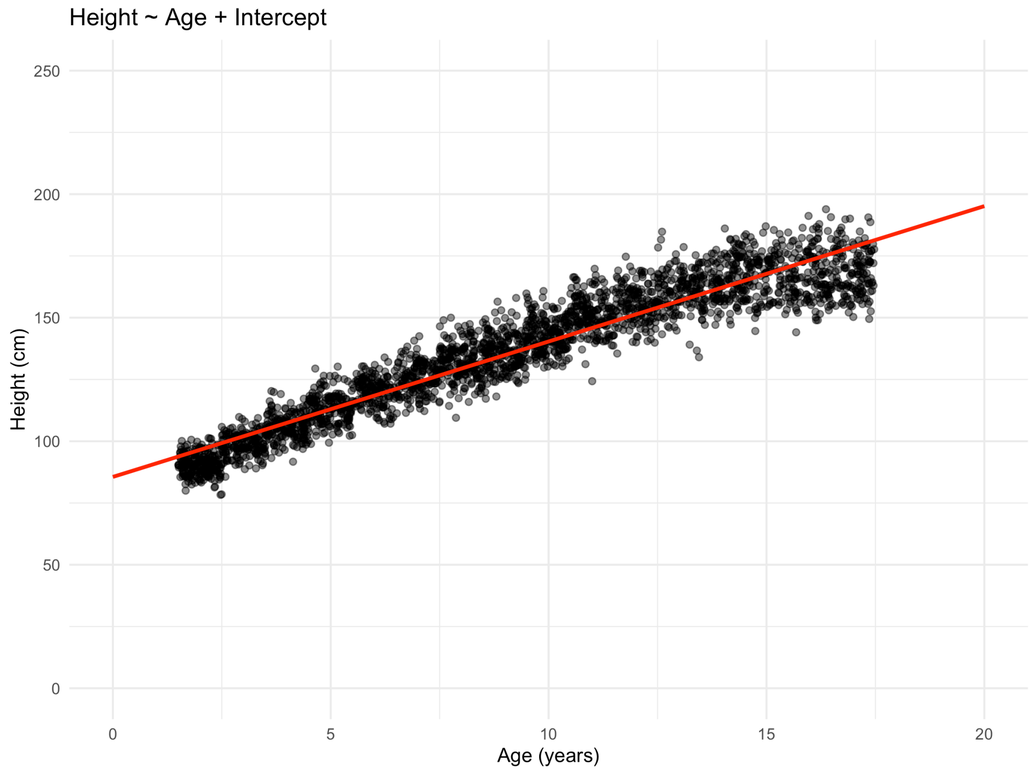



```
Call:
lm(formula = Height_cm ~ Age, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1814  -5.4882   0.0058   5.5853  30.1664

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56411    0.37412   228.7   <2e-16 ***
Age          5.48009    0.03645   150.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.456 on 2570 degrees of freedom
Multiple R-squared:  0.8979,    Adjusted R-squared:  0.8979
F-statistic: 2.26e+04 on 1 and 2570 DF,  p-value: < 2.2e-16
```

Linear Models in R

Statistical model

- Model 2 - model height as a function of age, with intercept

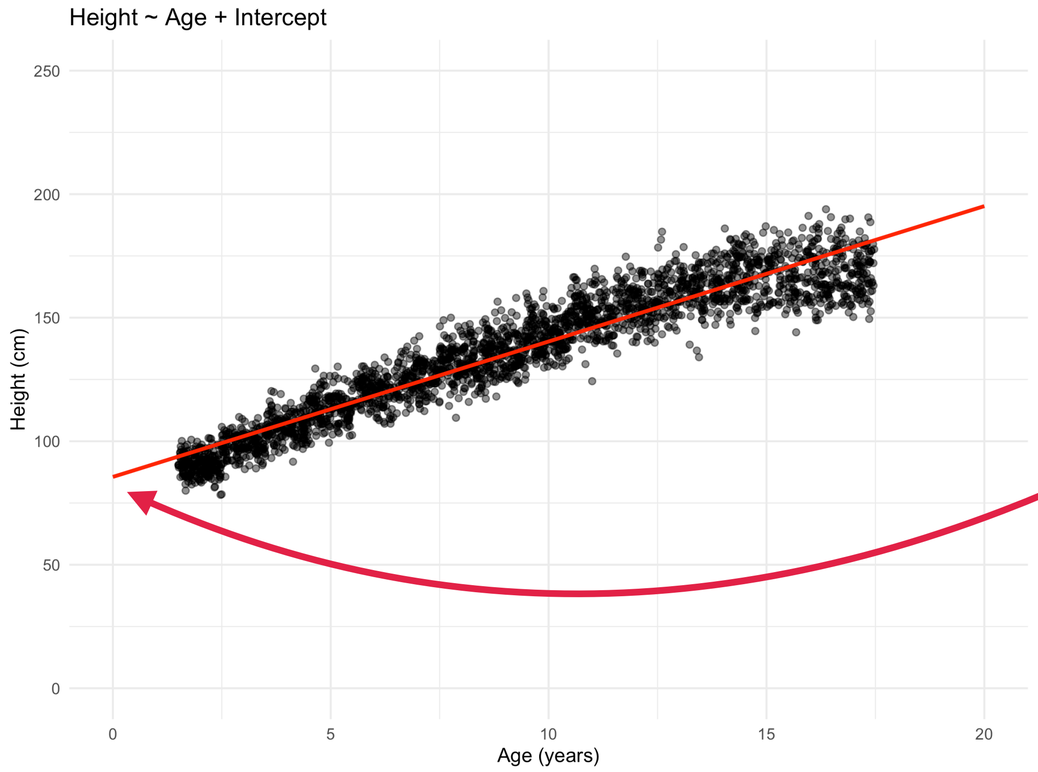



```
Call:
lm(formula = Height_cm ~ Age, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1814  -5.4882   0.0058   5.5853  30.1664

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56411    0.37412   228.7   <2e-16 ***
Age          5.48009    0.03645   150.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.456 on 2570 degrees of freedom
Multiple R-squared:  0.8979,    Adjusted R-squared:  0.8979
F-statistic: 2.26e+04 on 1 and 2570 DF,  p-value: < 2.2e-16
```

Linear Models in R

Statistical model

- Model 2 - model height as a function of age, with intercept

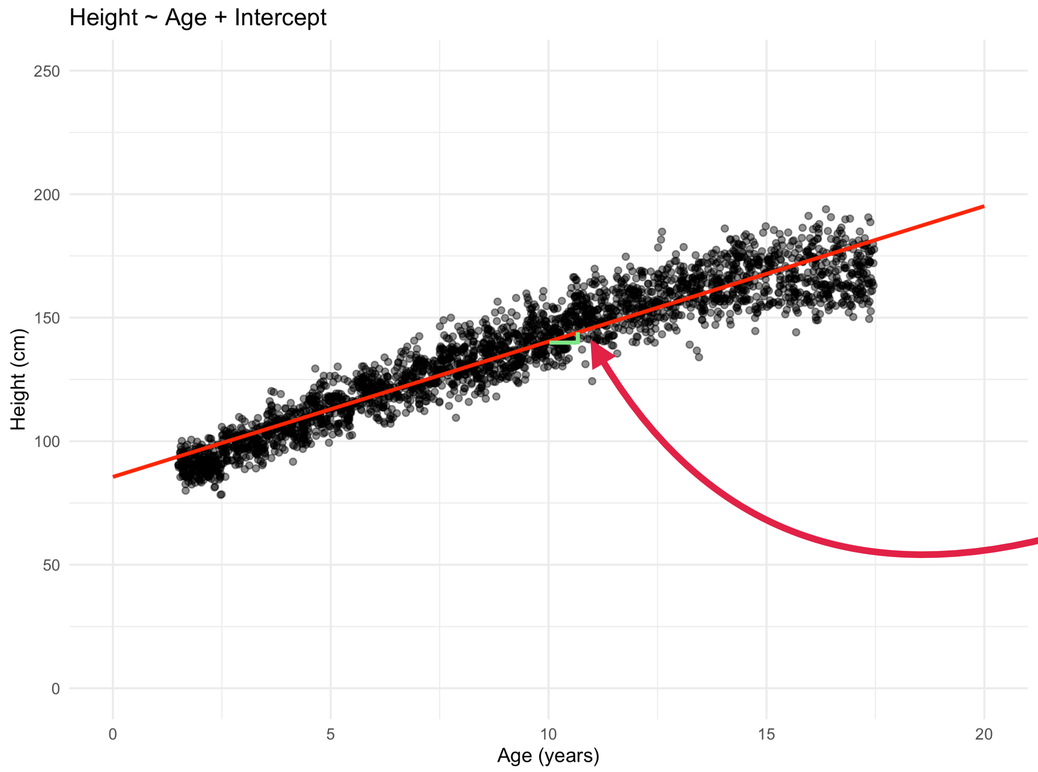



```
Call:
lm(formula = Height_cm ~ Age, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1814  -5.4882   0.0058   5.5853  30.1664

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56411    0.37412   228.7   <2e-16 ***
Age          5.48009    0.03645   150.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.456 on 2570 degrees of freedom
Multiple R-squared:  0.8979,    Adjusted R-squared:  0.8979
F-statistic: 2.26e+04 on 1 and 2570 DF,  p-value: < 2.2e-16
```

Linear Models in R

Statistical model

- Model 2 - model height as a function of age, with intercept

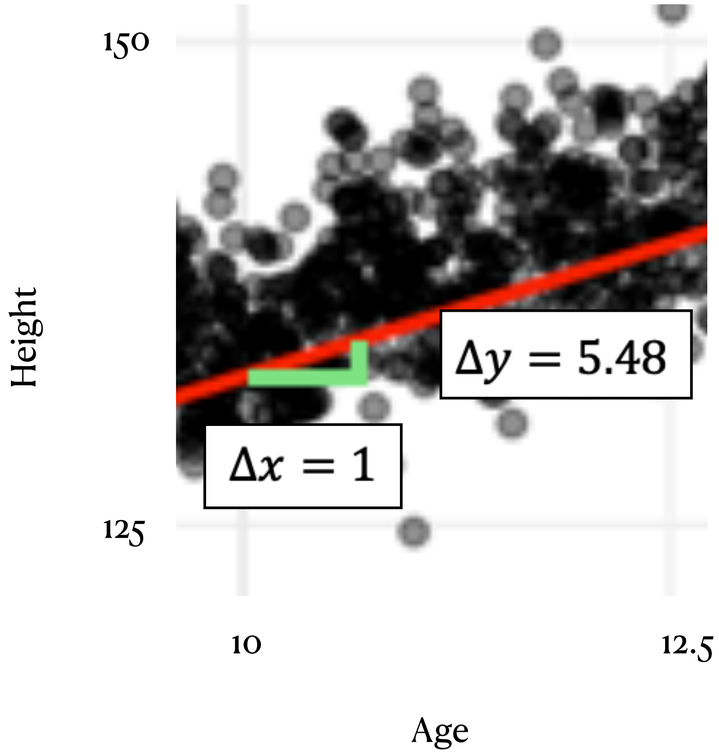



```
Call:
lm(formula = Height_cm ~ Age, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1814  -5.4882   0.0058   5.5853  30.1664

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56411    0.37412   228.7   <2e-16 ***
Age          5.48009    0.03645   150.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.456 on 2570 degrees of freedom
Multiple R-squared:  0.8979,    Adjusted R-squared:  0.8979
F-statistic: 2.26e+04 on 1 and 2570 DF,  p-value: < 2.2e-16
```

Linear Models in R

Statistical model

- Model 2 - model height as a function of age, with intercept




```
Call:
lm(formula = Height_cm ~ Age, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1814  -5.4882   0.0058   5.5853  30.1664

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56411    0.37412   228.7   <2e-16 ***
Age          5.48009    0.03645   150.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.456 on 2570 degrees of freedom
Multiple R-squared:  0.8979,    Adjusted R-squared:  0.8979
F-statistic: 2.26e+04 on 1 and 2570 DF,  p-value: < 2.2e-16
```

Linear Models in R

Statistical model

- Do My Predictors Matter? Statistical Significance of Regression Coefficients

$$error_i = y_i - \hat{y}_i$$

$$SS_{error} = \sum_{i=n}^{n} (y_i - \hat{y}_i)^2$$

$$MS_{error} = \frac{SS_{error}}{df} = \frac{SS_{error}}{N - p}$$

$$SE_{model} = \sqrt{MS_{error}}$$

$$SE_{\hat{\beta}_x} = \frac{SE_{model}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$t_{N-p} = \frac{\hat{\beta}}{SE_{\hat{\beta}_x}}$$

```
Call:
lm(formula = Height_cm ~ Age, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1814  -5.4882   0.0058   5.5853  30.1664

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56411    0.37412   228.7   <2e-16 ***
Age          5.48009    0.03645   150.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.456 on 2570 degrees of freedom
Multiple R-squared:  0.8979,    Adjusted R-squared:  0.8979
F-statistic: 2.26e+04 on 1 and 2570 DF,  p-value: < 2.2e-16
```

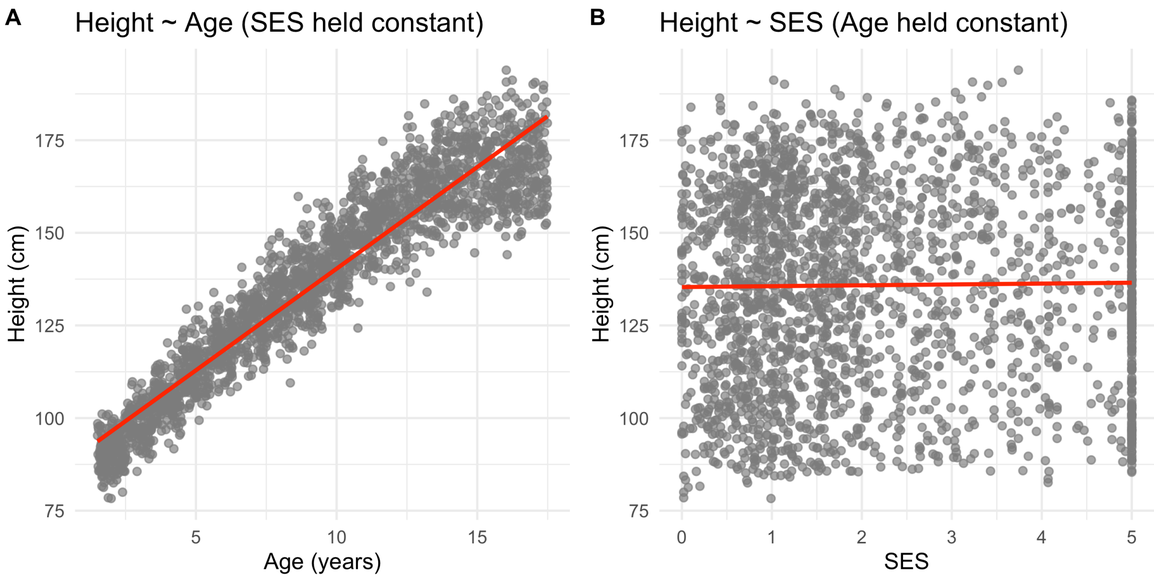
Linear Models in R

Statistical model

- Model 3 - multiple linear regression model

$$Height_i = \hat{\beta}_0 + \hat{\beta}_1 * Age_i + \hat{\beta}_2 * SES_i + \epsilon$$

```
Call:
lm(formula = Height_cm ~ Age + SES, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-29.5644  -5.4616   0.0534   5.3712  27.4854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.97657    0.45647 186.160   <2e-16 ***
Age          5.48464    0.03849 142.482   <2e-16 ***
SES          0.23219    0.11403   2.036   0.0418 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.488 on 2307 degrees of freedom
Multiple R-squared:  0.8983,    Adjusted R-squared:  0.8982
F-statistic: 1.018e+04 on 2 and 2307 DF,  p-value: < 2.2e-16
```



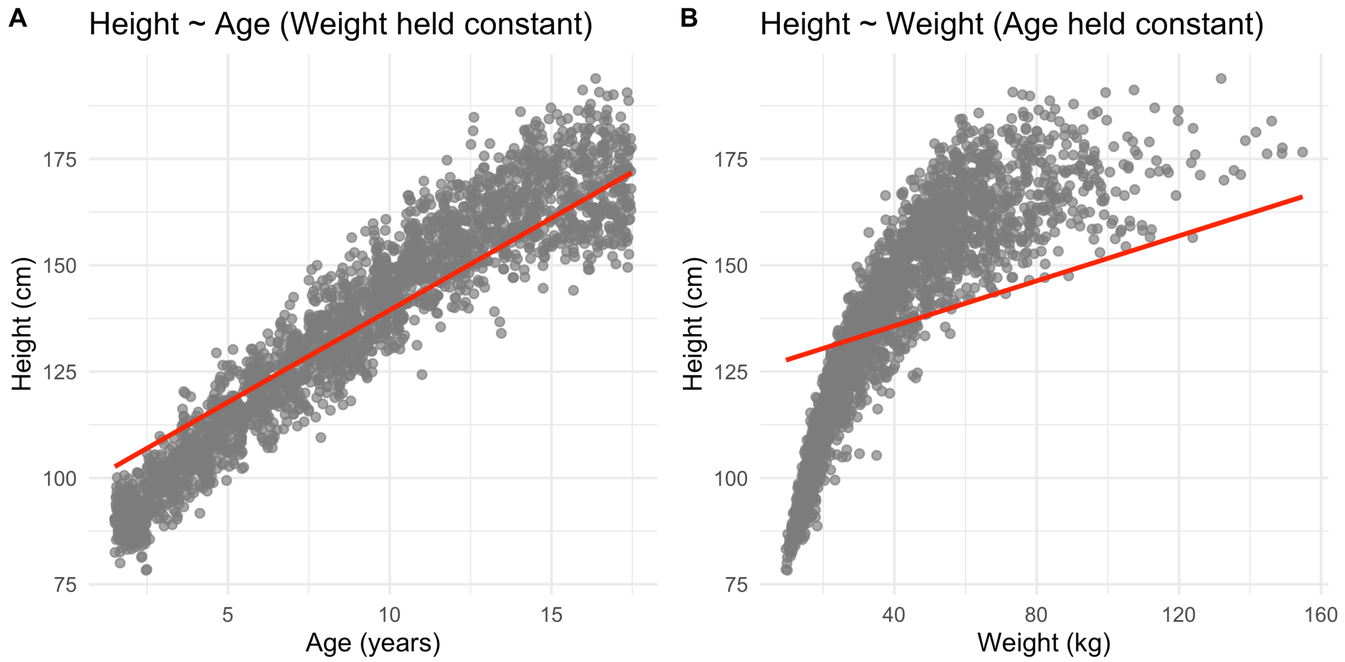
Linear Models in R

Statistical model

- Model 4 - multiple linear regression model

$$Height_i = \hat{\beta}_0 + \hat{\beta}_1 * Age_i + \hat{\beta}_2 * Weight_i + \epsilon$$

```
Call:
lm(formula = Height_cm ~ Age + Weight_kg, data = df_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-36.817  -4.784   0.199   5.240  25.309

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.39187    0.34163  249.96   <2e-16 ***
Age          4.32865    0.06008   72.05   <2e-16 ***
Weight_kg    0.26439    0.01148   23.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.713 on 2568 degrees of freedom
Multiple R-squared:  0.915,     Adjusted R-squared:  0.915
F-statistic: 1.383e+04 on 2 and 2568 DF,  p-value: < 2.2e-16
```



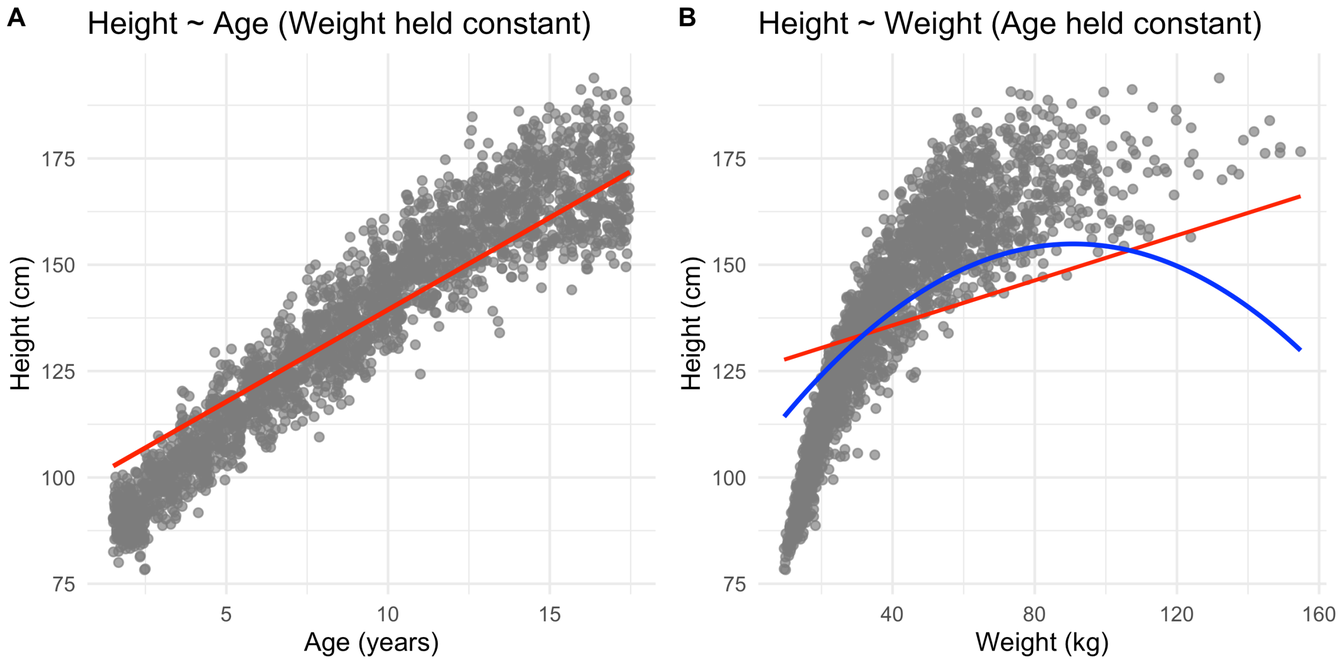
Linear Models in R

Statistical model

- Model 5 - multiple linear regression model with quadratic term

$$Height_i = \hat{\beta}_0 + \hat{\beta}_1 * Age_i + \hat{\beta}_2 * Weight_i + \hat{\beta}_3 * (Weight_i)^2 + \epsilon$$

```
Call:
lm(formula = Height_cm ~ Age + Weight_kg + I(Weight_kg^2), data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-27.2362  -4.1028   0.0229   4.3074  28.7013

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    75.6922626  0.4420063  171.25   <2e-16 ***
Age             3.1063900  0.0664443   46.75   <2e-16 ***
Weight_kg       1.1146395  0.0304907   36.56   <2e-16 ***
I(Weight_kg^2) -0.0061316  0.0002079  -29.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.667 on 2567 degrees of freedom
Multiple R-squared:  0.9365,    Adjusted R-squared:  0.9365
F-statistic: 1.263e+04 on 3 and 2567 DF,  p-value: < 2.2e-16
```



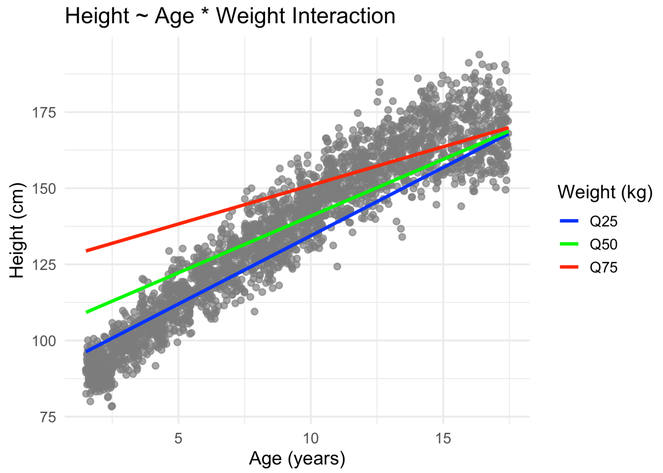
Linear Models in R

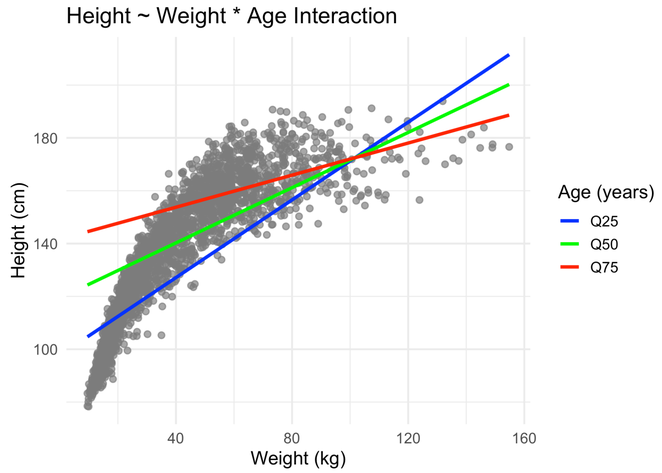
Statistical model

- Model 6 - multiple linear regression model with interaction terms

$$Height_i = \hat{\beta}_0 + \hat{\beta}_1 * Age_i + \hat{\beta}_2 * Weight_i + \hat{\beta}_3 * (Age_i * Weight_i) + \epsilon$$

```
Call:
lm(formula = Height_cm ~ Age * Weight_kg, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-20.7934  -4.0082  -0.0459   3.7624  22.0740

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   68.105378   0.537755  126.65   <2e-16 ***
Age            5.643966   0.059785   94.41   <2e-16 ***
Weight_kg      1.026475   0.022373   45.88   <2e-16 ***
Age:Weight_kg -0.055500   0.001484  -37.40   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.207 on 2567 degrees of freedom
Multiple R-squared:  0.945,     Adjusted R-squared:  0.9449
F-statistic: 1.471e+04 on 3 and 2567 DF,  p-value: < 2.2e-16
```





Linear Models in R

Statistical model

- How well is the model describing the data? Coefficient of Determination.

$$SS_{total} = SS_{model} + SS_{error}$$

$$\sum_{i=n}^{n}(y_i - \bar{y})^2 = \sum_{i=n}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=n}^{n}(y_i - \hat{y}_i)^2$$

Variance = variance_explained + variance_not_explained

$$R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$$

It indicates that about **89.79% of the variance in Height_cm** is explained by **Age** in your linear model.

```
Call:
lm(formula = Height_cm ~ Age, data = df_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1814  -5.4882   0.0058   5.5853  30.1664

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56411    0.37412   228.7   <2e-16 ***
Age          5.48009    0.03645   150.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.456 on 2570 degrees of freedom
Multiple R-squared:  0.8979,    Adjusted R-squared:  0.8979
F-statistic: 2.26e+04 on 1 and 2570 DF,  p-value: < 2.2e-16
```

Hands-on

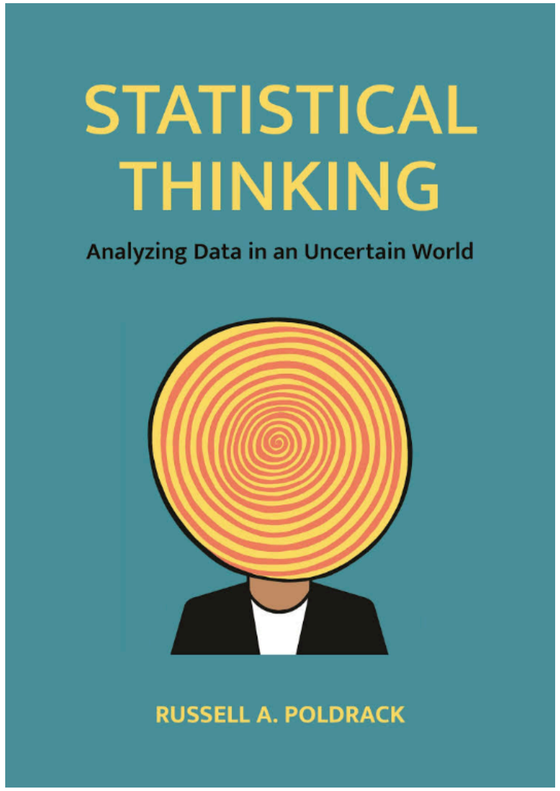
github.com/asantoangles/linear_models_in_R

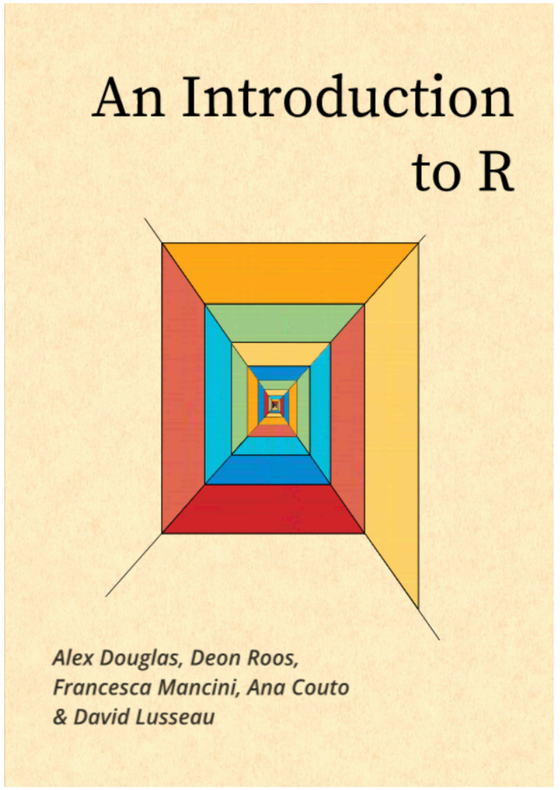
script_02_linear_models.R

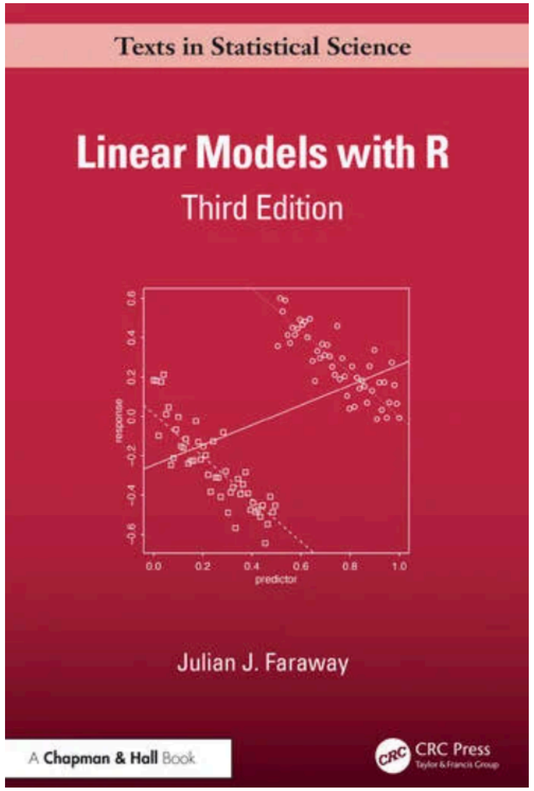
Resources

https://bookdown.org/palmjulia/r_intro_script/

https://intro2r.com/







Thanks!

github.com/asantoangles