Quantitative Methoden 1

Tutorium 09/04/2021

- Exploratory data analysis
- Linear models
- R exercise- create a model

Exploratory data analysis
Basic terms review

- **Mean**: the sum of the observed values divided by the number of observations
- **Median**: the number in the middle
- **Variance** and **Standard deviation** : The variance is the average squared distance from the mean. The standard deviation is the square root of the variance.
- **Interquartile range(IQR)** The IQR interquartile range is the length of the box in a box plot. It is computed as IQR=Q3−Q1, where Q1 and Q3 are the 25th and 75th percentiles, respectively.
- **Quantile** : measures of the position of a distribution. A p-quantile indicates at which value of a distribution p% of the values lie below this value.
- **Quartile** are the quantiles at 25%, 50% and 75%. The quartile at 25% is also referred to as Q1 ("lower quartile"), that at 50% as Q2 ("middle quartile" = "median") and that at 75% as Q3 ("upper quartile"). They are one of the most frequently used form of quantiles in statistics.
- **Percentile** divide a distribution into 100 equal parts, i.e. into 1% parts, and are therefore percentages. For example, the percentile P40 corresponds to the point in the distribution below which 40% of all values in a distribution lie

❗The standard deviation is useful when considering how far the data are distributed from the mean.

https://openintro-ims.netlify.app/summarizing-visualizing-data.html#categorical-data
http://www.abschlussarbeit.at/index.php/sitemap.html?id=65

Chapter Review Exercise IMS book 2.5.2-1
Group discussion

Make-up exam. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

1. Does the new student's score increase or decrease the average score?
2. What is the new average?
3. Does the new student's score increase or decrease the standard deviation of the scores?

Mean. The sample mean can be calculated as the sum of the observed values divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The standard deviation represents the typical deviation of observations from the mean. Often about 68% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, these percentages are not strict rules.

https://openintro-ims.netlify.app/summarizing-visualizing-data.html#chp2-review-exercises

Chapter Review Exercise 1
Group discussion

Make-up exam. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an **average score of 74** points with a **standard deviation of 8.9** points. The student who took the make-up the following day scored **64 points** on the exam.

1. Does the new student's score increase or decrease the average score?
2. What is the new average?
3. Does the new student's score increase or decrease the standard deviation of the scores?

Mean. The sample mean can be calculated as the sum of the observed values divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The standard deviation represents the typical deviation of observations from the mean. Often about 68% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, these percentages are not strict rules.

1. Decrease
2. (74*24+64)/25=73.6
3. Increase, The concept of standard deviation give us an idea of the range of the scores and is represented as (mean ± standard deviation) = [mean - standard deviation ; mean + standard deviation]. In this case we have (74 ± 8.9) = [65.1; 82.9]. This implies that all scores are included in this range. So if we add a score that it is not included in the range the standard deviation will increase.

https://openintro-ims.netlify.app/summarizing-visualizing-data.html#chp2-review-exercises

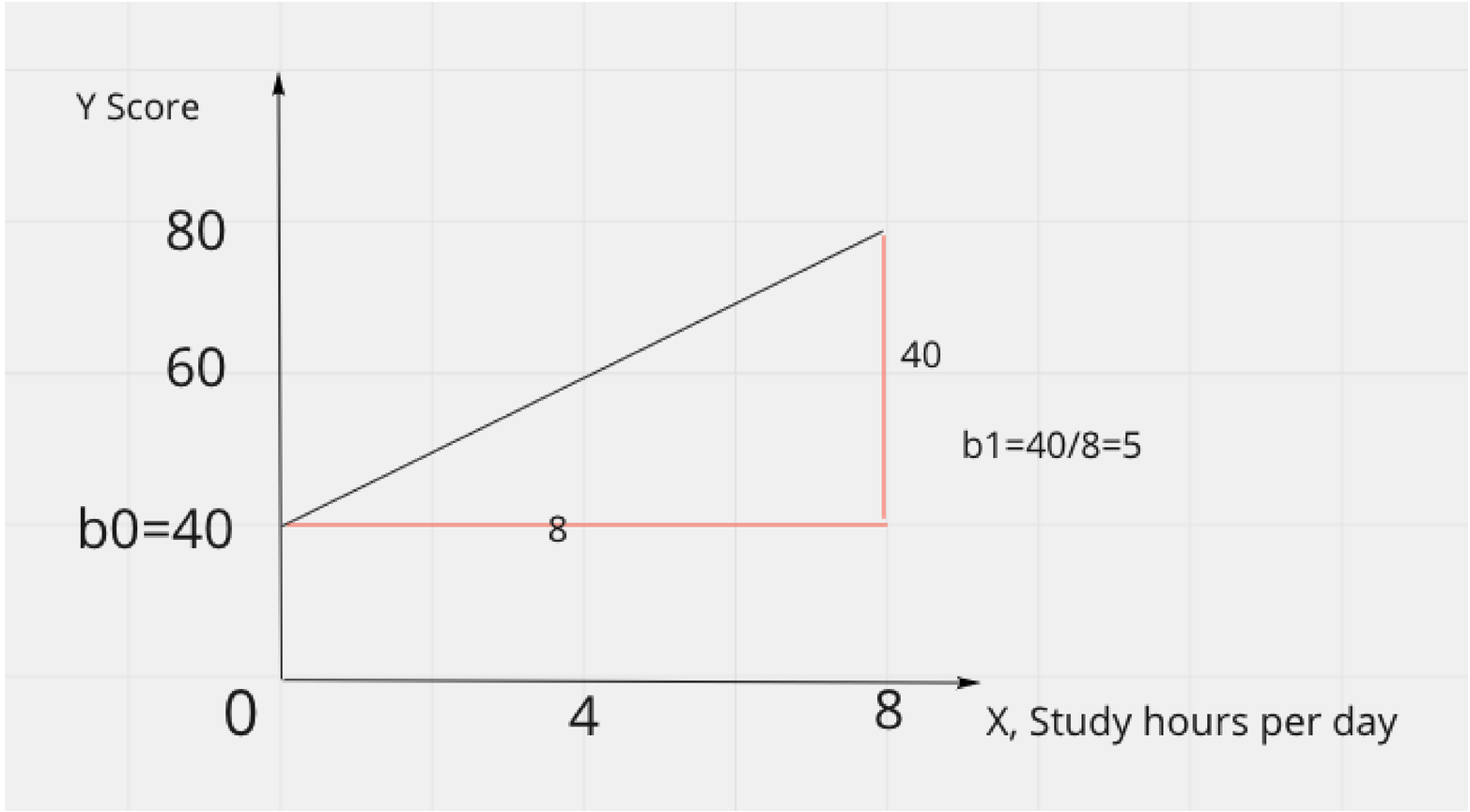
Linear models

- Linear regression is the statistical method for fitting a line to data where the relationship between two variables x and y, can be modeled by a straight line with some error,
- When we use x to predict y, we usually call x the **predictor** variable and we call y the **outcome**

Y = b0 + b1 x
b0: intercept, b1:slope

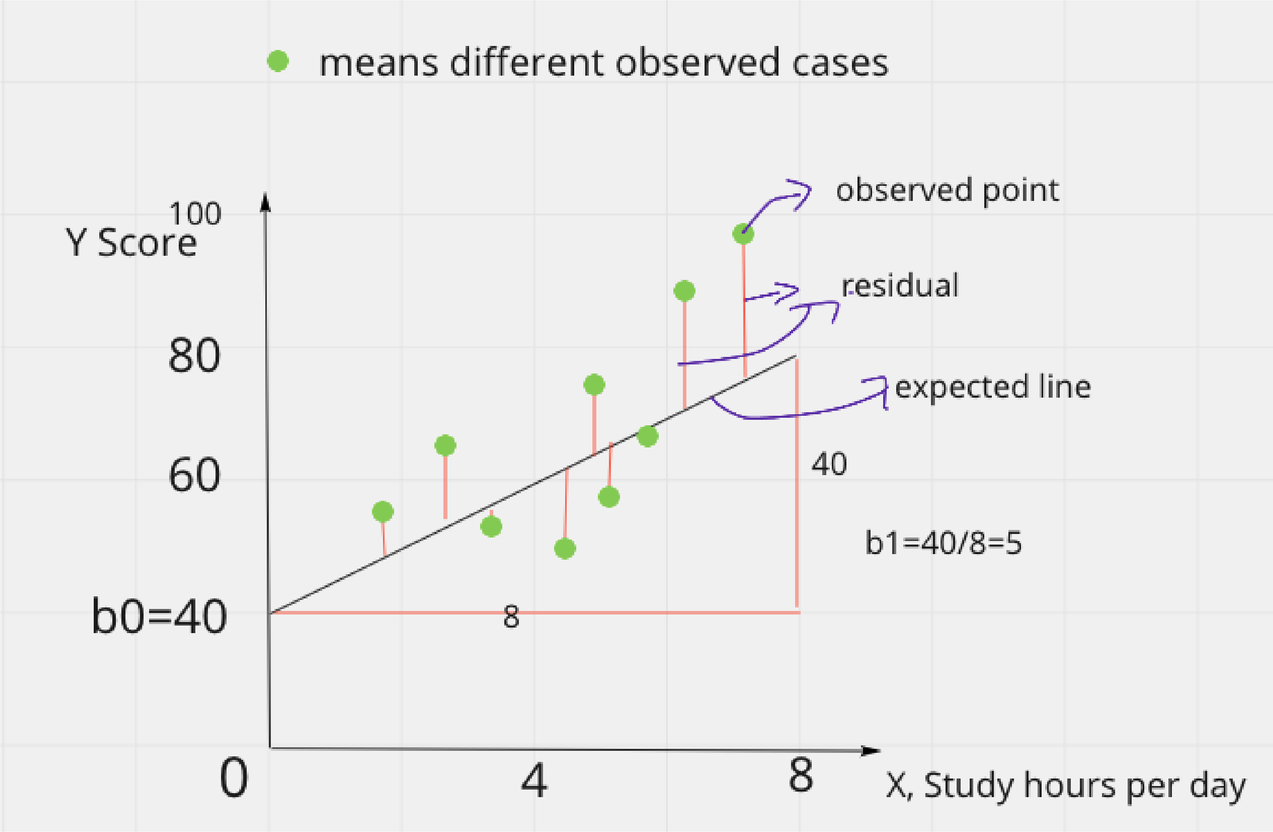
Y = 40 + 5 x

The steepness of a hill is called a **slope**



https://openintro-ims.netlify.app/summarizing-visualizing-data.html#chp2-review-exercises

Linear models
Residual



Residual: Difference between observed and expected. The residual of the i^{th} observation (x_i, y_i) is the difference of the observed outcome (y_i) and the outcome we would predict based on the model fit $(\hat{y}_i)$:

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

We typically identify $\hat{y}_i$ by plugging x_i into the model.

https://openintro-ims.netlify.app/summarizing-visualizing-data.html#chp2-review-exercises

Linear models
Least squares line

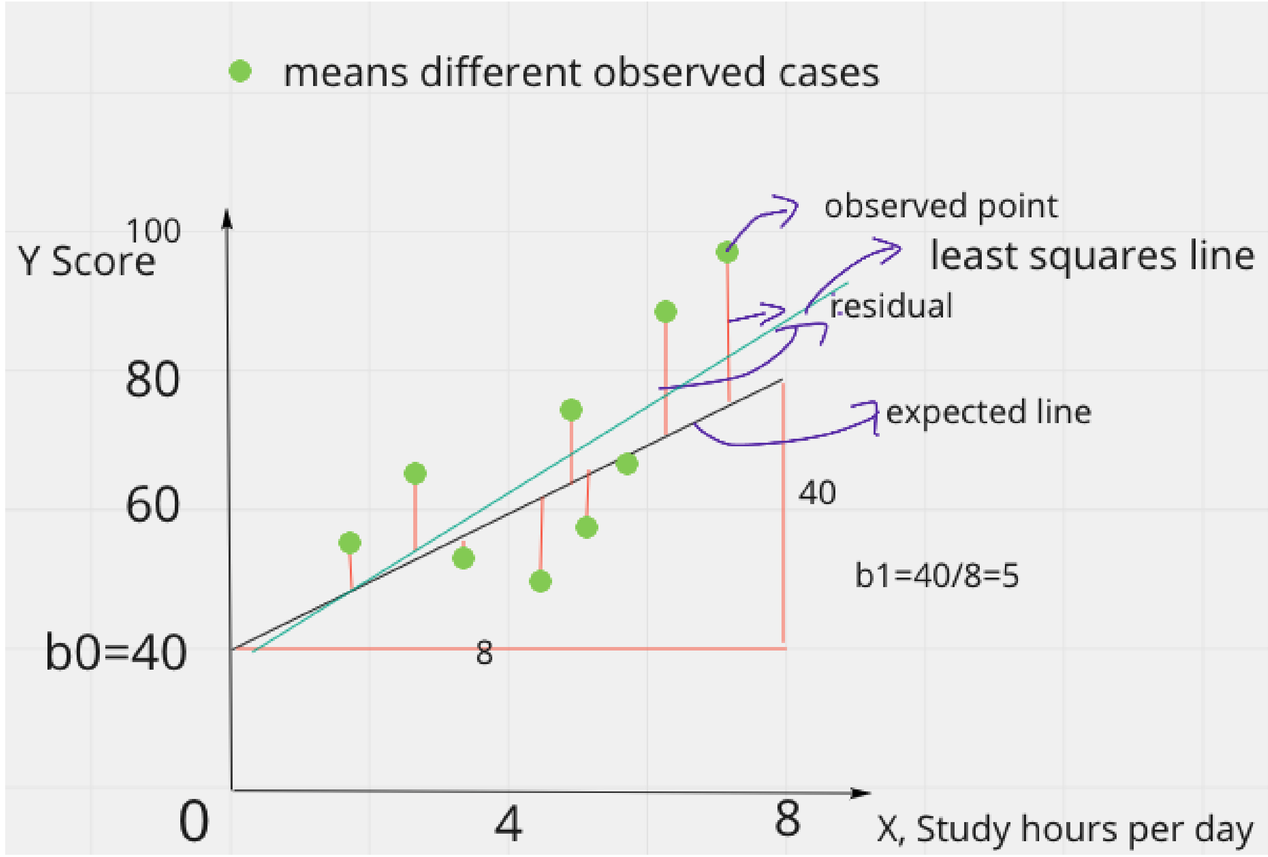
We want a line that has small residuals that means a line represented by observations has the least gap with the expected line...

Least squares line let us better understand what R squared is

However, a more common practice is to choose the line that minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

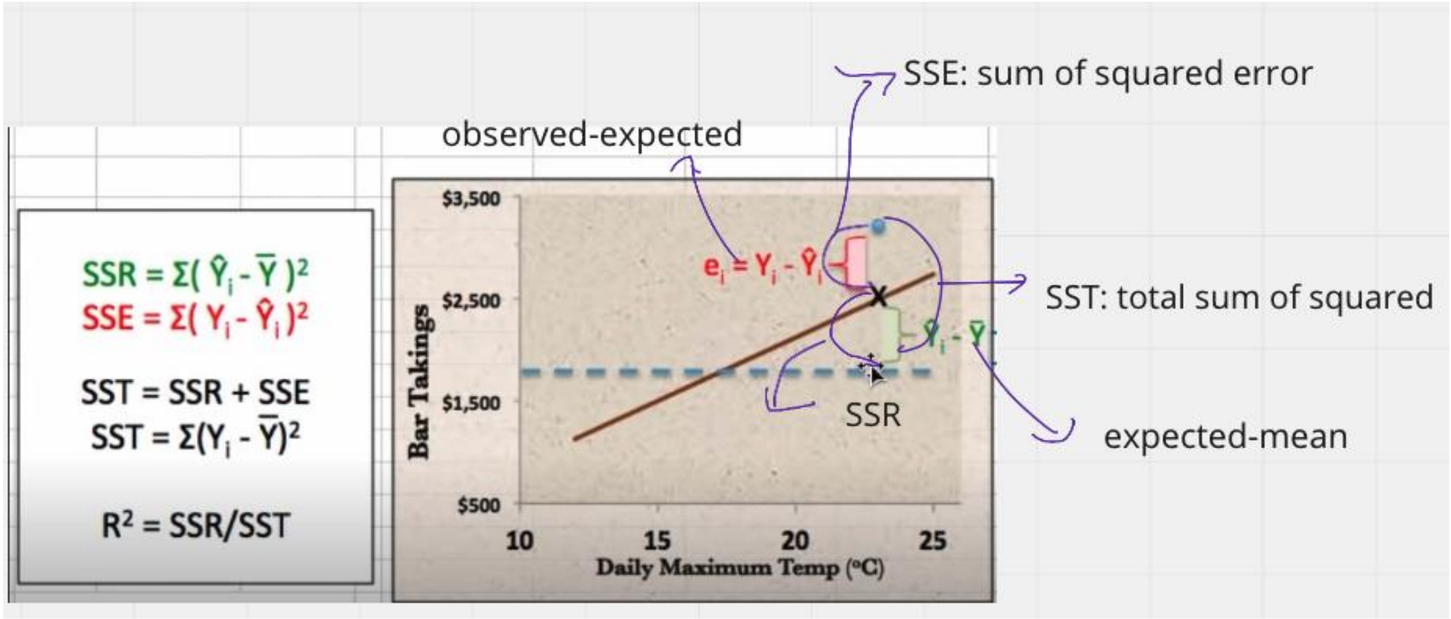
y=b0+b1x



https://openintro-ims.netlify.app/summarizing-visualizing-data.html#chp2-review-exercises

Linear models
R-squared: describe the strength of a fit

The R squared describes the amount of variation in the outcome variable that is explained by the least squares line



Choose a reference object or benchmark, here we set the line that is the average of the outcome, where has no trend.

The best model is
The distance from the observation point to the average is equal to the distance from the expected value to the average, d.h. the observation point is on the expected value, at this time r-squared is equal to 1

❗0 < R squared <1

- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

❗We tend to choose the model with bigger R squared

https://www.youtube.com/watch?v=aq8VU5KLmkY&t=740s
https://openintro-ims.netlify.app/summarizing-visualizing-data.html#chp2-review-exercises

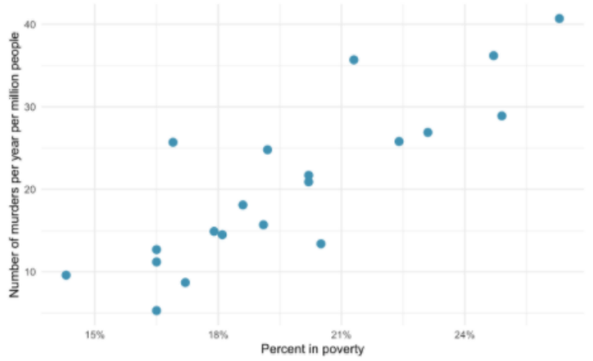
R Exercise IMS Book 3.2.7/9

Murders and poverty, regression. The following regression output is for predicting annual murders per million (annual_murders_per_mil) from percentage living in poverty (perc_pov) in a random sample of 20 metropolitan area

- Write out the linear model.
- Interpret the intercept.
- Interpret the slope.
- Interpret R2.
- Calculate the correlation coefficient.

term	estimate	std.error	statistic	p.value
(Intercept)	-29.90	7.79	-3.84	0.001
perc_pov	2.56	0.39	6.56	0.000

s	R-squared	Adjusted R-squared
5.51	70.52%	68.89%



R Exercise IMS Book 3.2.7/9

1. Data download

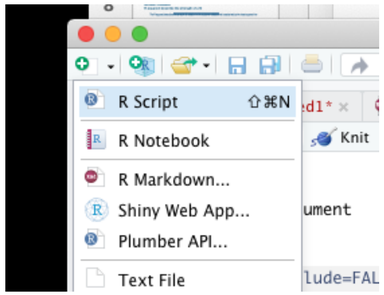
1. https://www.openintro.org/data/index.php?data=murders

We do not have provenance for these d

Downloads

- CSV file
- Tab-delimited text file
- R source
- R Data file

2 Establish a new rmd.datei and Save workplace (the same place where you save the csv file)




```
[Workspace loaded from ~/Documents/.RData]

> setwd("~/Desktop")
>
```

R Exercise 3.2.7/9

3. Read the data and Establish the model, at the beginning, you could copy the following code and execute it, to see what will happen (#means the comment, don't need copy)

```
#read the data
murders <- read.csv("murders.csv")

# have a look of the data
library(dplyr)
glimpse(murders)


head (murders)
```

```
```{r}
#lm() linear model function, so here y is annual_murders_per_mil, x is perc_pov
m_murders_poverty <- lm(murders$annual_murders_per_mil ~ murders$perc_pov)

#summary() let r do the calculation
summary(m_murders_poverty)
```
```

R Exercise 3.2.7/9

4. Interpret the outcome

```
Call:
lm(formula = murders$annual_murders_per_mil ~ murders$perc_pov)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1663 -2.5613 -0.9552  2.8887 12.3475

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -29.901      7.789  -3.839   0.0012 **
murders$perc_pov   2.559      0.390   6.562 3.64e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.512 on 18 degrees of freedom
Multiple R-squared:  0.7052,    Adjusted R-squared:  0.6889
F-statistic: 43.06 on 1 and 18 DF,  p-value: 3.638e-06
```

Answer:

(a) Write out the linear model.

Since we are given a regression output, the value for β0 and β1 are provided by the first column titled “Estimate” respectively.

$\bar{y} = -29.901 + 2.559 \cdot x$

(b) Write out the Intercept.

Intercept/ b0: -29.901

(c) Write out the slope.

slope/ b1: 2.559

(d) R-squard: 0.7052

(e) Calculate the correlation coefficient:

Since we know R2.

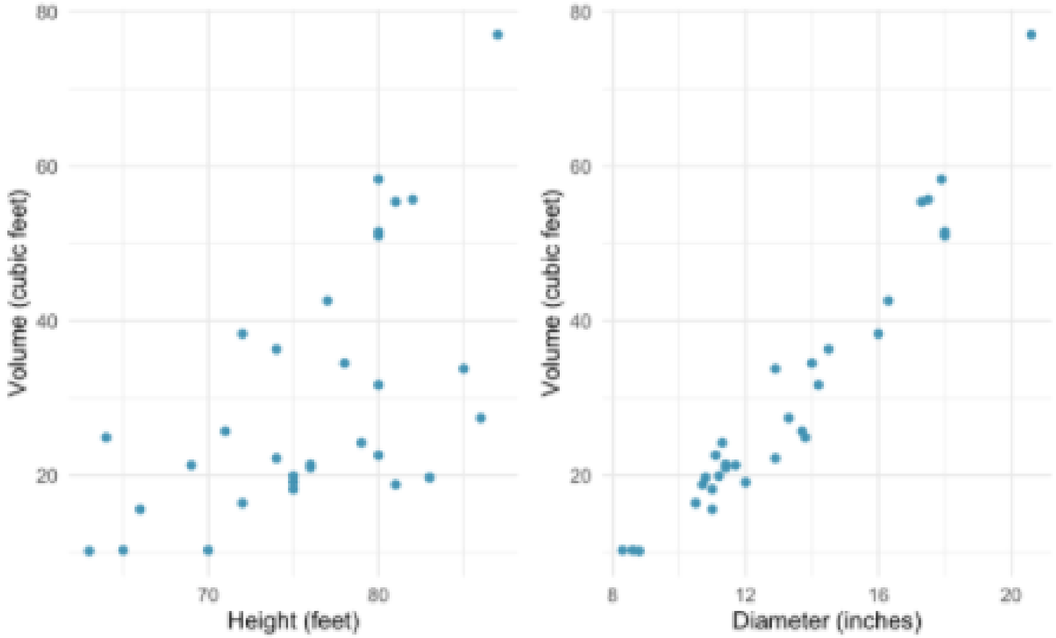
The correlation coefficient is simply the square root of R2

http://rstudio-pubs-static.s3.amazonaws.com/328229_f8a33dd9d02c4f26b78adaa0cc0956af.html

Chapter review exercise

- **Trees.** The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground

1. Describe the relationship between volume and height of these trees.
2. Describe the relationship between volume and diameter of these trees.
3. Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

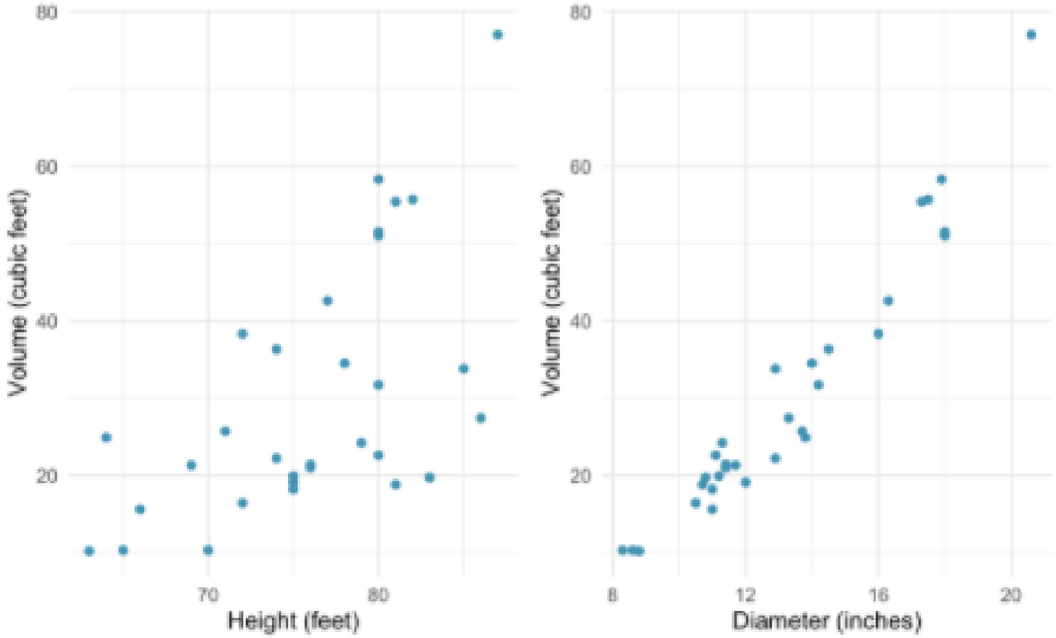


https://openintro-ims.netlify.app/intro-linear-models.html#chp3-review

Chapter review exercise

- **Trees.** The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground

1. Describe the relationship between volume and height of these trees.
2. Describe the relationship between volume and diameter of these trees.
3. Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.



1. The first plot shows an upward trend that, while evident, is not as strong as the second.
2. The second plot shows a relatively strong upnward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y.
3. Diameter. Because Height seems not has a clear linear relationship with volume.

https://openintro-ims.netlify.app/intro-linear-models.html#chp3-review