Simple Linear Regression

April 6, 2021

Plan

Here's what to expect today

1. Determine relationship between X and Y
    - $y = mx + b$
    - What is the "best" line?
    - Slope and intercept interpretation
2. Rules and assumptions
    - What can we do with model?
    - What can't we do?
3. Residuals
    - Use to check model diagnostics
    - Verify assumptions

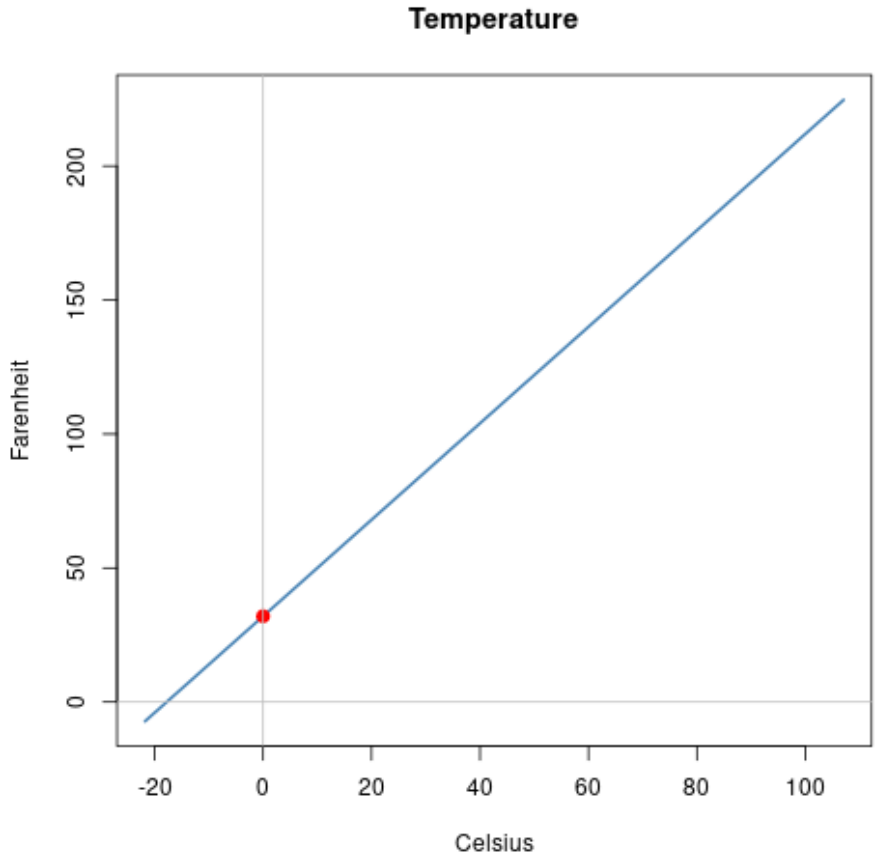
Introduction

Previously in the course, we considered *scatterplots*, depicting the relationship between two continuous variables, along with the *correlation*, indicating the linear relationship between the two, with $|r| \leq 1$

Today we will consider defining a functional linear relationship between two continuous variables, known as simple linear regression.

Will will designate these variables Y, the *response or dependent* variable, and X, the *explanatory or independent* variable. Although this suggests a causal relationship, we will limit ourselves to interpreting it as a functional relationship for now

Linear Function (temperature)



$$F = 1.8C + 32$$

Questions

Typically, we step are attempting to answer the following questions with the use of simple linear regression:

- Is there a linear relationship between X and Y?
- How does one variable change in response to a change in another?
- If an association exists, can it be exploited to predict the value of one variable, based on the assumed value of another?

Examples

- X = an individual's metacarpal bone length
  Y = individual's height

- X = time a person with high cholesterol is on the drug Lipitor
  Y = change in person's cholesterol level over time

- X = nitrate emissions rate from industries in a region
  Y = nitrate concentration for rainfall in the region

Simple Linear Regression

As the name suggests, simple linear regression involves describing the relationship between X and Y with a straight line of the form

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 indicates the *intercept* for the line, and β_1 the *slope*

As we should rarely expect that our data will fit perfectly on a straight line, we account for this error with the ϵ term, taken to be a random variable

The usual assumption we have is $\epsilon \sim N(0, \sigma^2)$. In particular, note that $E(\epsilon) = 0$, so that

$$\begin{aligned} E(Y) &= E(\beta_0 + \beta_1 X + \epsilon) \\ &= E(\beta_0 + \beta_1 X) + E(\epsilon) \\ &= \beta_0 + \beta_1 X \end{aligned}$$

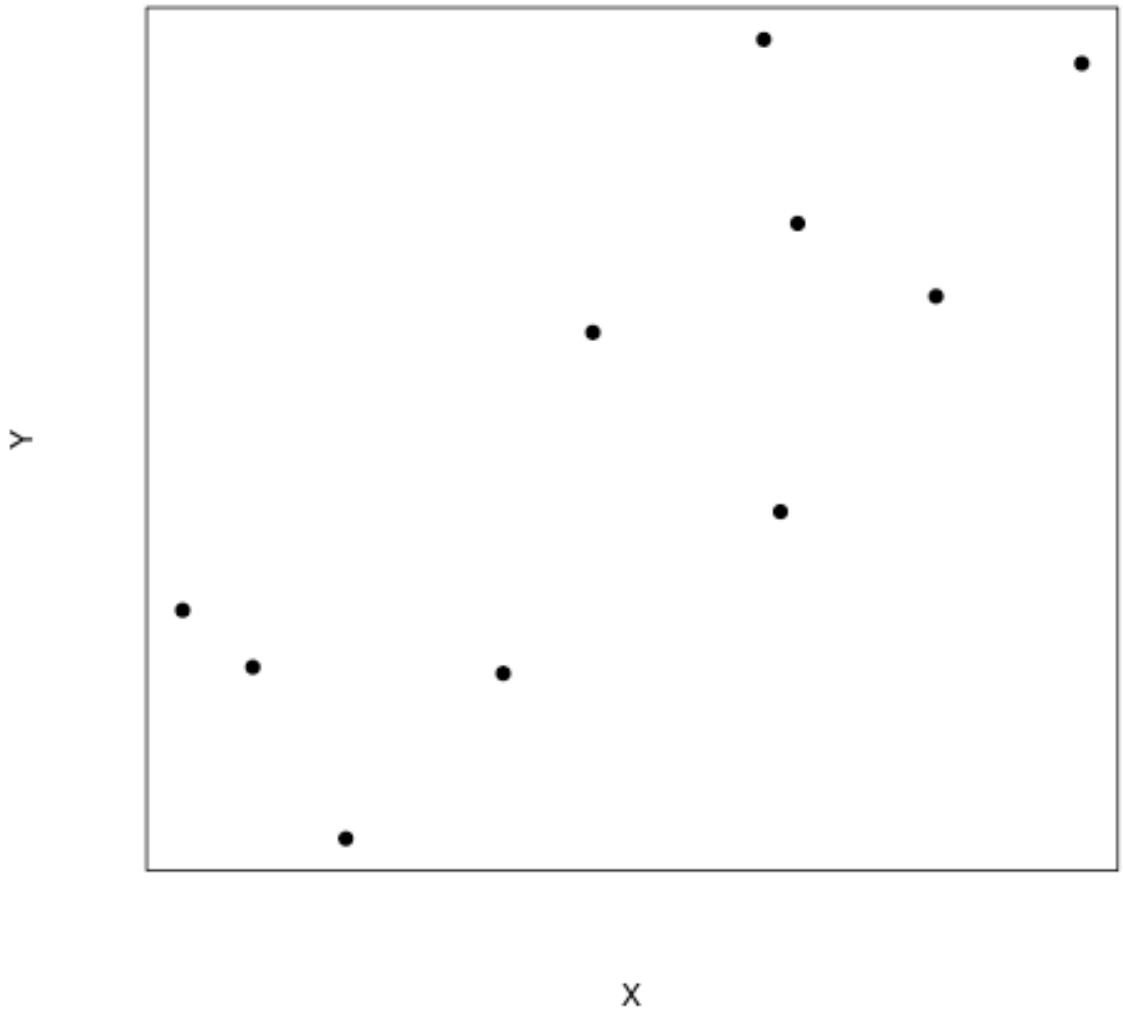
Determining a line

We will start with our n observations, represented by pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, for which a scatterplot can be generated

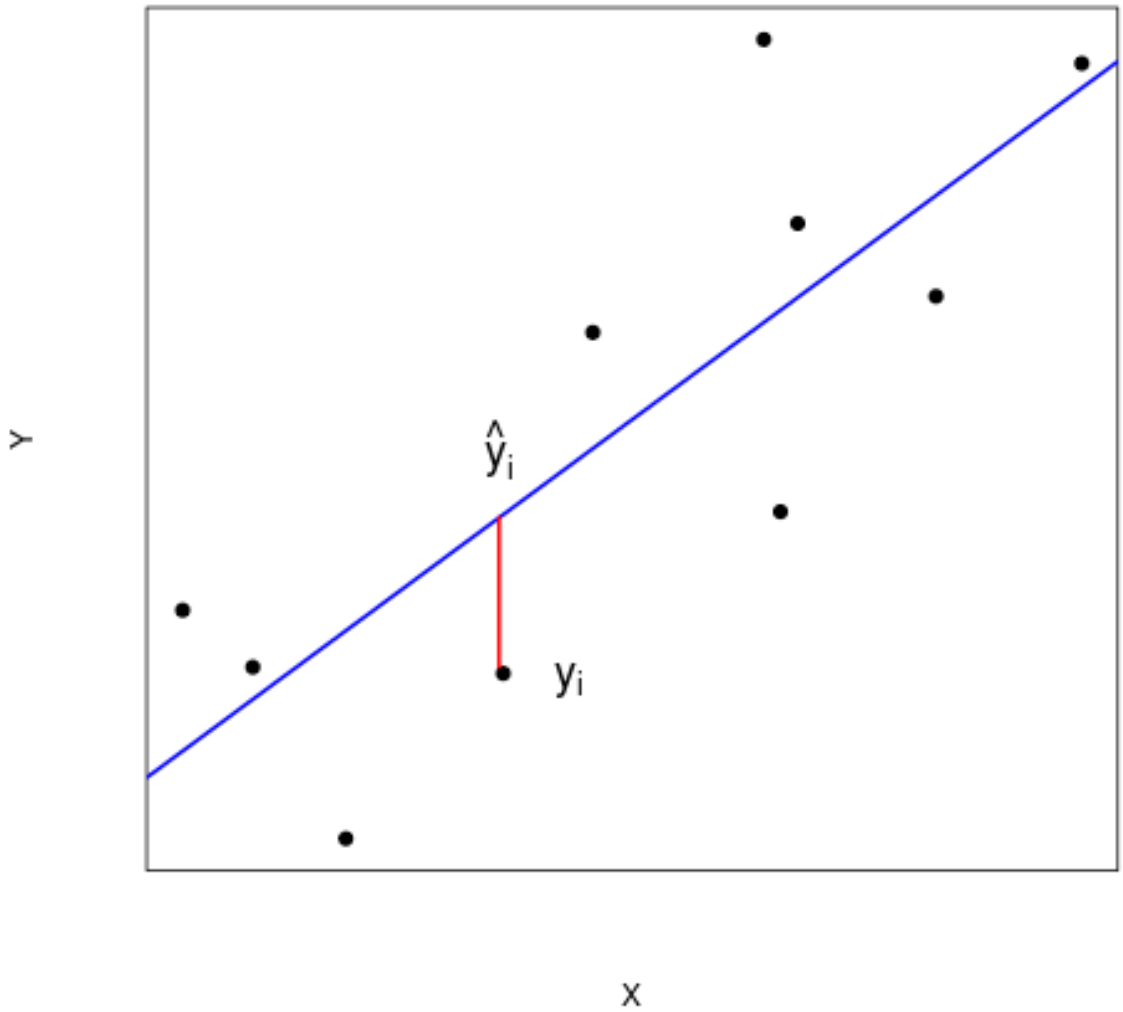
For each data point (x_i, y_i), let $\hat{y}_i$ denote the corresponding point of our (currently hypothetical) *best* line

An estimate of the error term we mentioned previously can be described as $e_i = (\hat{y}_i - y_i)$, representing the vertical distance between our observed point and the corresponding point on the line. This is known as our *residual*

Determining a line



Determining a line



Least-Squares Method

The *least-squares* method for determing our line is the one in which the quantity $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ is minimized, where

$$\hat{Y} = b_0 + b_1 X$$

where $b_1 = r\left(\frac{s_y}{s_x}\right)$ and $b_0 = \overline{Y} - b_1\overline{X}$, where $r = corr(X, Y)$, and s_x and s_y are the standard deviations for X and Y, respectively

From this, we note two things:

1. The intercept term, b_0, is understood as the mean value of Y when $X = 0$
2. The slope of the line, b_1, represents the change in Y for a one-unit increase in X

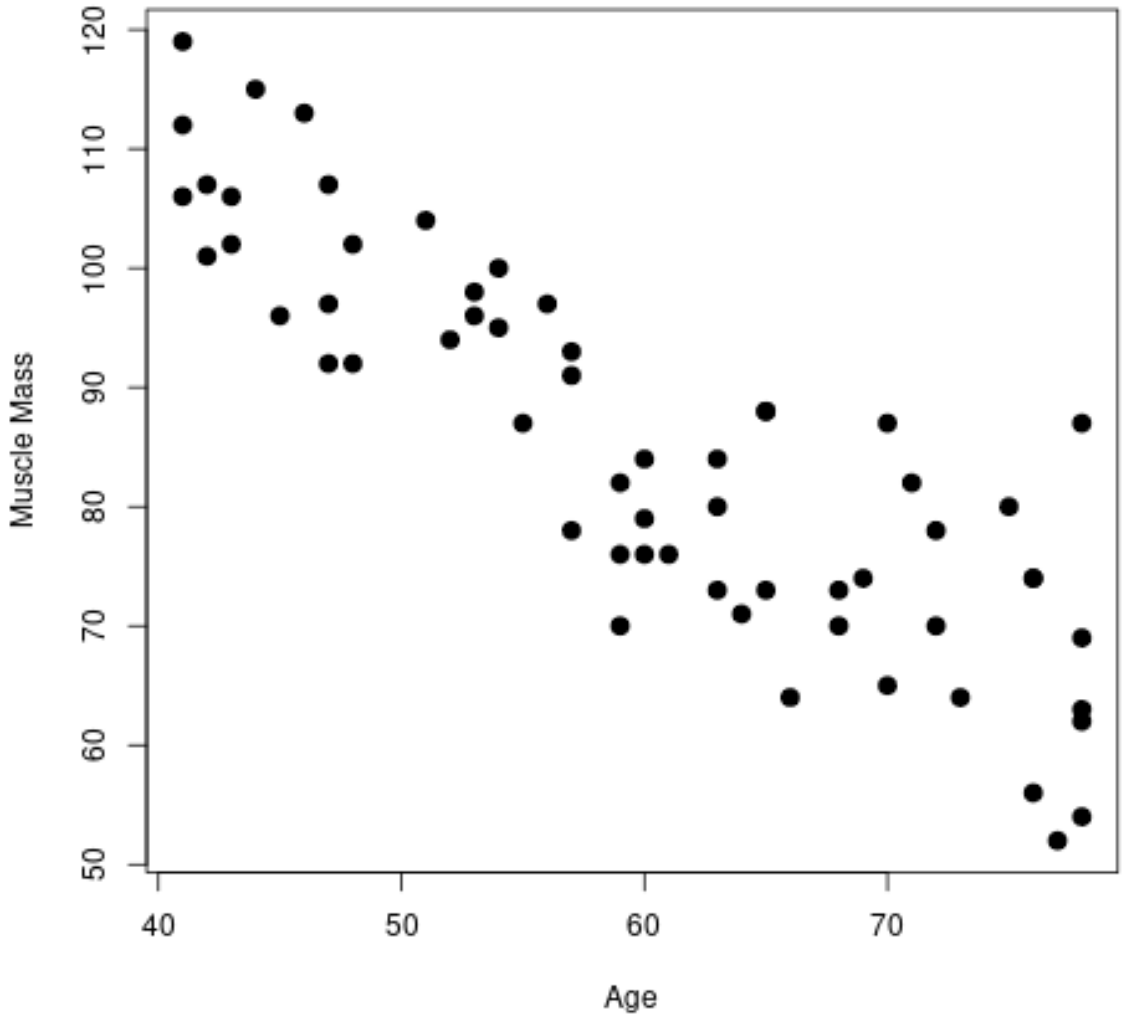
Example - Muscle Mass

It is expected that an individual's muscle mass will descrease with age

To investigate this relationship in women, a nutritionist randomly sampled 15 women from each 10 year age group, beginning with age 40 and ending with age 79

We will start by considering a scatter plot of the data

Example - Muscle Mass



Example - Muscle Mass

Summary Statistics:

$$\overline{x} = 59.983, \quad s_x = 11.979$$
$$\overline{y} = 84.967, \quad s_y = 16.209$$

$$r = -0.866$$

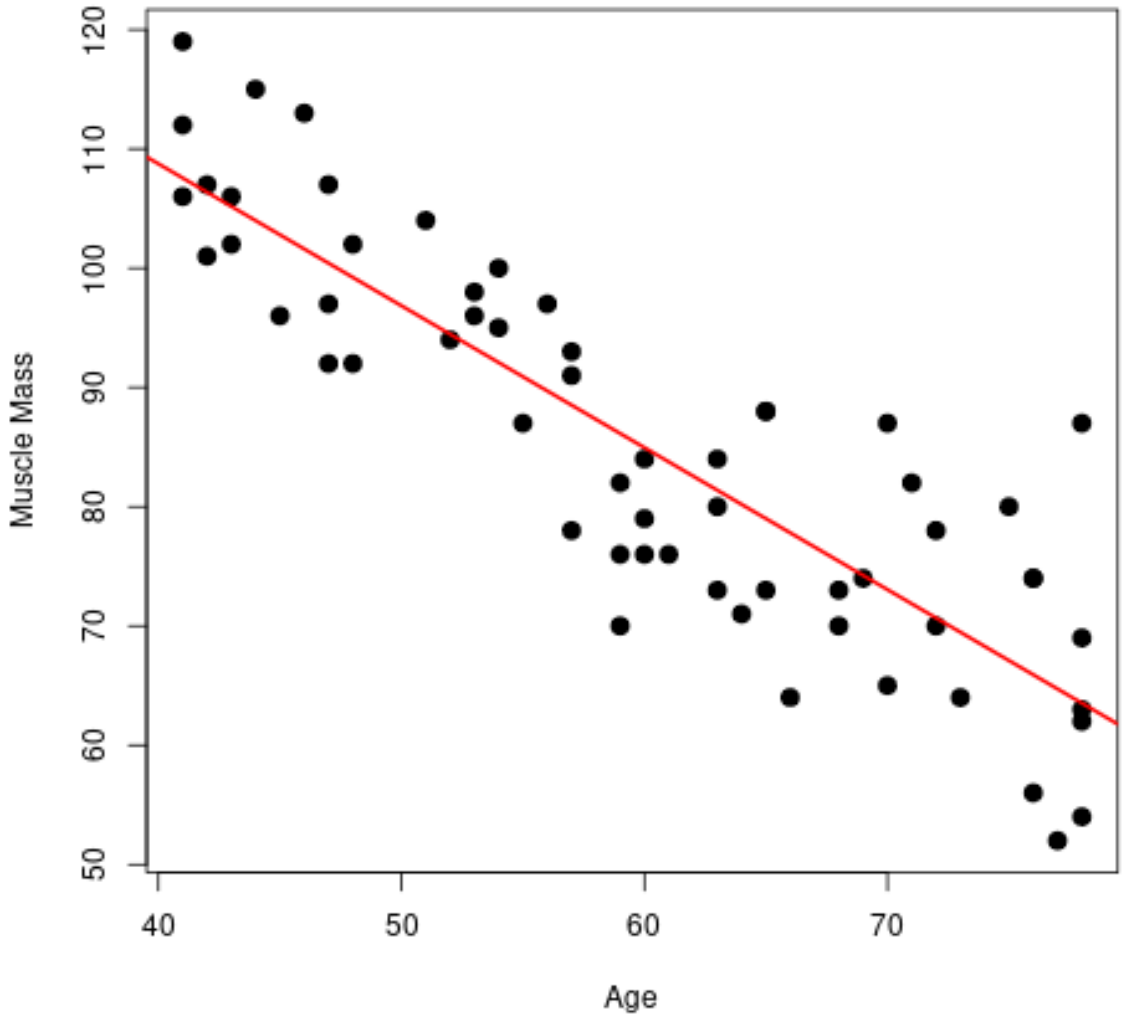
The least squares line is given by $\hat{y} = b_0 + b_1 x$, where

$$b_1 = r\left(\frac{s_y}{s_x}\right) = -0.866\left(\frac{16.209}{11.979}\right) = -1.19$$

and

$$b_0 = \overline{y} - b_1\overline{x} = 84.967 - (-1.19) \times 59.983 = 156.35$$

Example - Muscle Mass



Example - Muscle Mass

$$\hat{y} = 156.35 - 1.19x$$

Conclusions:

1. For every one year increase in age (x), we would expect an decrease of about 1.2 pounds in muscle mass
2. When $x = 0$, the average muscle mass of a woman is expected to be around 156 lbs - *Note that this doesn't make any sense*. We can only interpret values that fall within the range of our model (40-80yrs)
3. For a woman who is 50 years old, we could predict her muscle mass to be

$$\hat{y} = 156.35 - 1.19 \times 50 = 96.85 \text{ lbs}$$

Inference and Simple Linear Regression

In most applications, when fitting a line with regression, our goal is to perform some kind of inference about the underlying relationship in the population

Our assumed model is of the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

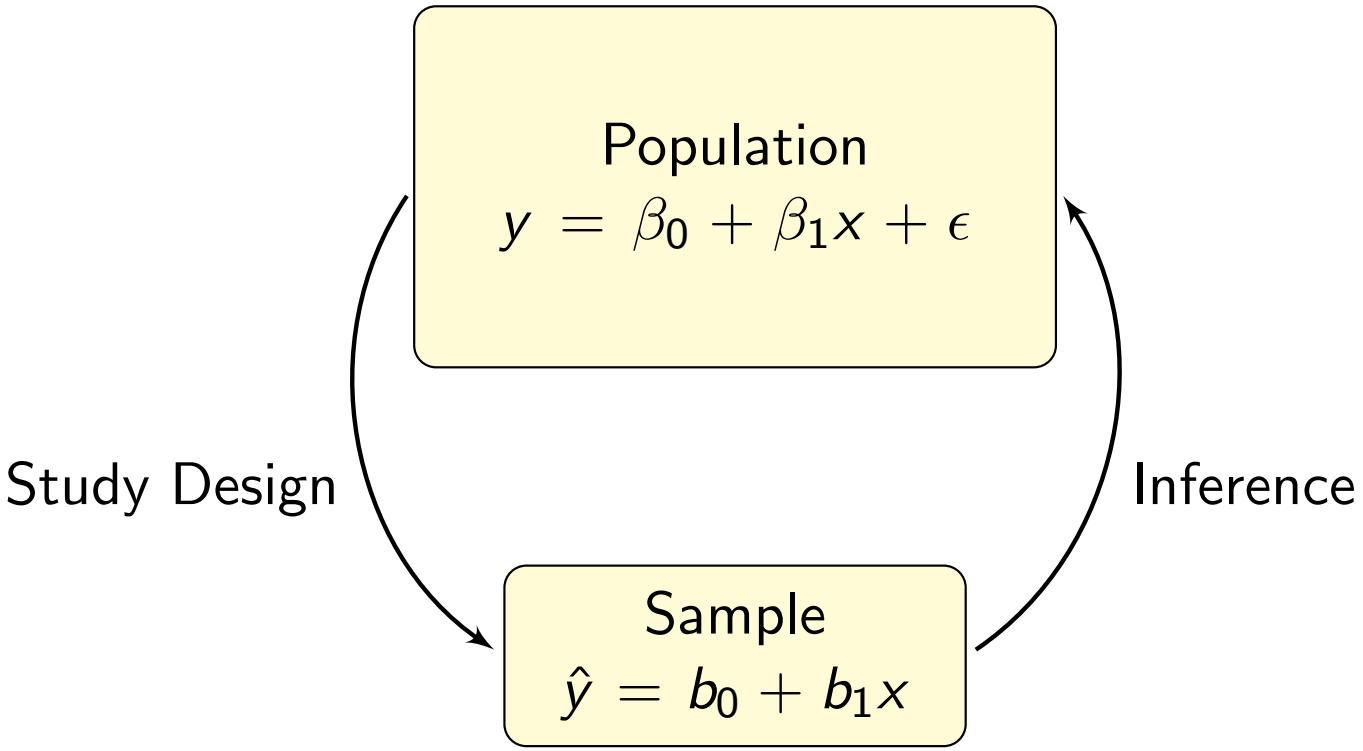
which is *approximated* with the least-squares line

$$\hat{y} = b_0 + b_1 x$$

The central limit theorem applies to the parameter estimates, with $b_i \sim N(\beta_i, var(\beta_i))$. From this, we can construct our t distribution, with

$$\frac{b - \beta}{sd(b)} \sim t_{n-1}$$

Inference and Simple Linear Regression



R

```
> fit <- lm(mass ~ age, data = muscle)
> summary(fit)

Call:
lm(formula = mass ~ age, data = muscle)

Residuals:
    Min      1Q  Median      3Q     Max
-16.137  -6.197  -0.597   6.761  23.473

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 156.3466     5.5123    28.4 <0.0000000000000002 ***
age          -1.1900     0.0902   -13.2 <0.0000000000000002 ***
---

Residual standard error: 8.17 on 58 degrees of freedom
Multiple R-squared:  0.75,      Adjusted R-squared:  0.746
F-statistic:  174 on 1 and 58 DF,  p-value: <0.0000000000000002
```

On scaling X

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

Critically, we see that the slope is intimately associated with the correlation between X and Y, but scaled with the ratio s_y/s_x to preserve the interpretation of units

A very common technique involves scaling the value of X, with

$$\tilde{x} = \frac{x - \overline{x}}{sd(x)}$$

so that $E(\tilde{x}) = 0$ $var(\tilde{x}) = 1$. With this, we have a new value for the slope,

$$\tilde{b}_1 = rs_y$$

indicating that a standard deviation change in X results in a $\tilde{b}_1$ change in Y

On scaling X

```
> fit <- lm(mass ~ scale(age, scale = TRUE), data = muscle)
> summary(fit)

Call:
lm(formula = mass ~ scale(age, scale = TRUE), data = muscle)

Residuals:
    Min      1Q  Median      3Q     Max
-16.137  -6.197  -0.597   6.761  23.473

Coefficients:
                          Estimate Std. Error t value             Pr(>|t|)
(Intercept)                  84.97       1.06    80.5 <0.0000000000000002
scale(age, scale = TRUE)    -14.04       1.06   -13.2 <0.0000000000000002

---

Residual standard error: 8.17 on 58 degrees of freedom
Multiple R-squared:  0.75,      Adjusted R-squared:  0.746
F-statistic:  174 on 1 and 58 DF,  p-value: <0.0000000000000002
```

R^2

A common metric used in regression, especially when considering predictive quality is the coefficient of determination, or R^2

Two values that are relevant here are the *total sum of squares* and the *residual sum of squares*,

$$SS_{Total} = \sum_{i=1}^{n}(y_i - \overline{y})^2, \qquad SS_{Residual} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

We then define R^2 as

$$R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

R^2

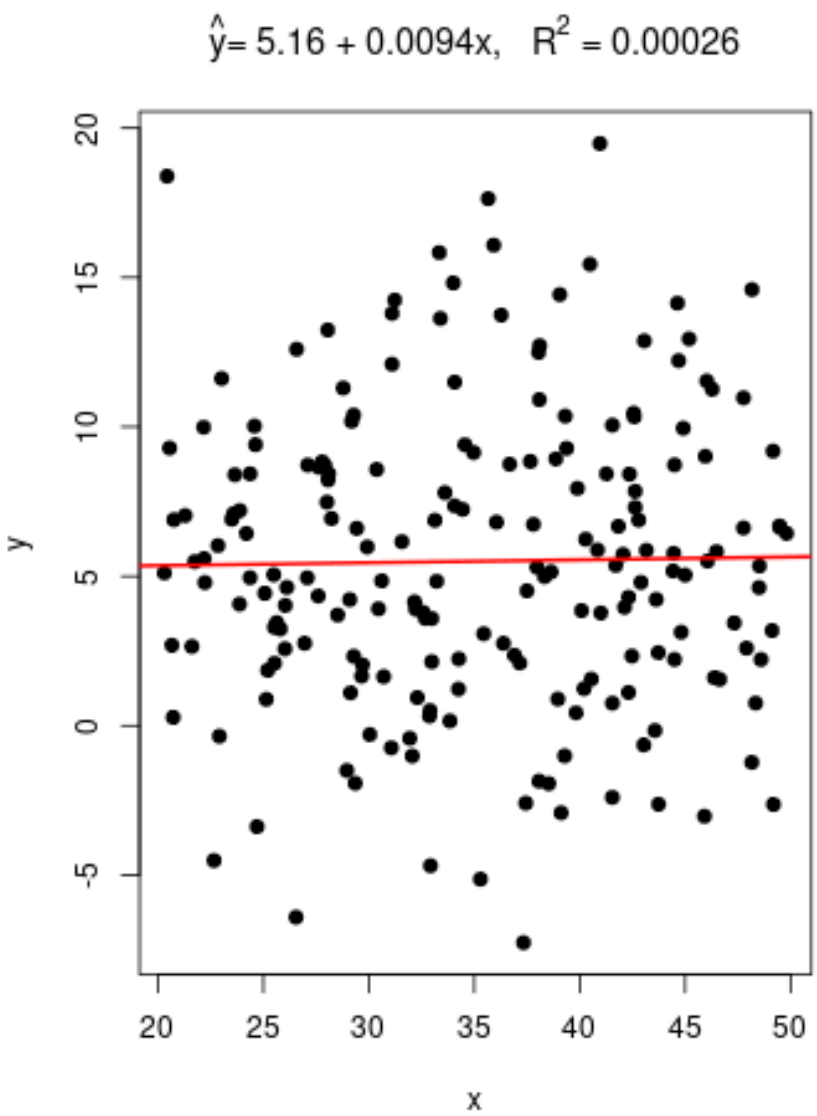
If the linear association (correlation) between x and y is small, b_1 will also be small, and $\hat{y} \approx \overline{y}$ and $SS_{Total} \approx SS_{Residual}$

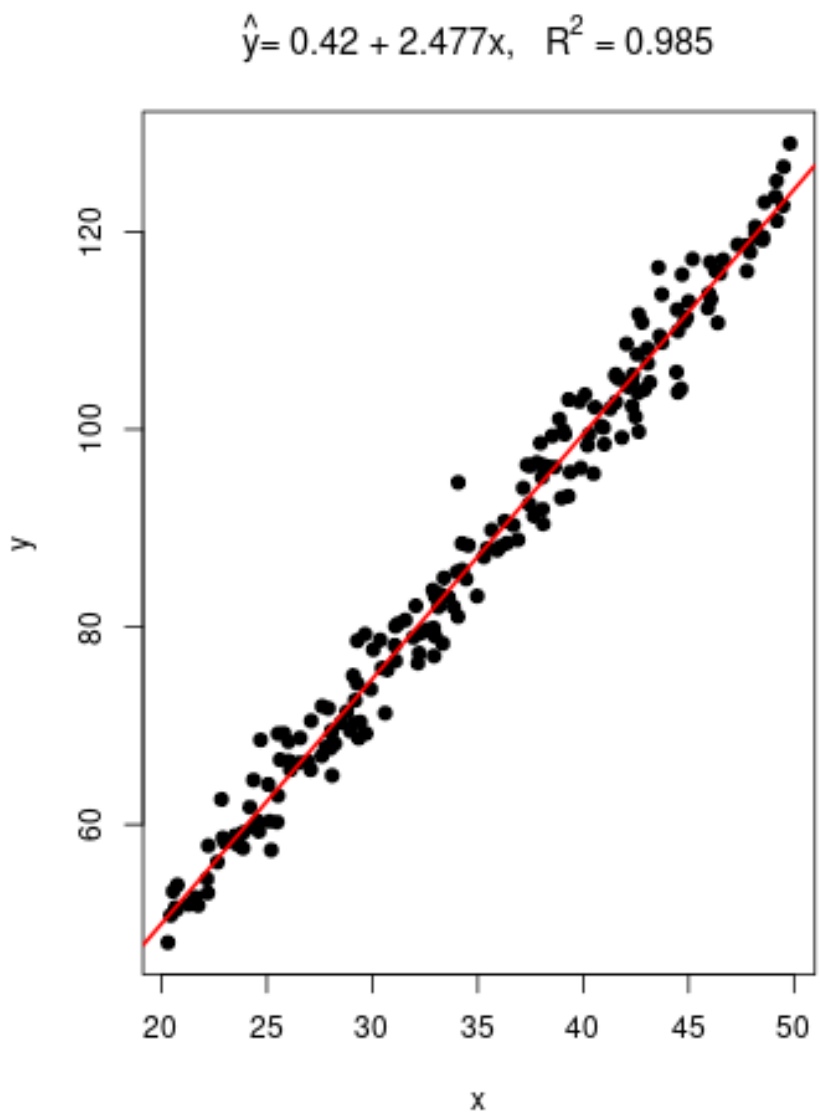
$$R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

$$SS_{Total} = \sum_{i=1}^{n}(y_i - \overline{y})^2, \qquad SS_{Residual} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\hat{y} = b_0 + b_1 x, \quad b_1 = r\left(\frac{s_y}{s_x}\right), \quad b_0 = \overline{y} - b_1\overline{x}$$

R^2


ŷ= 5.16 + 0.0094x, R² = 0.00026
y
x
20
15
10
5
0
-5
20
25
30
35
40
45
50


ŷ= 0.42 + 2.477x, R² = 0.985
y
x
120
100
80
60
20
25
30
35
40
45
50

Residuals

As we noted previously, our assumed model has the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

where we assumed that $\epsilon \sim N(0, \sigma^2)$. For each observation fit with the model, we have an estimate of this error term, known as the *residual*,

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \hat{y}_i$$

The validity of our linear model is built on a collection of assumptions, which can be investigated by examining the set of model residuals e_i

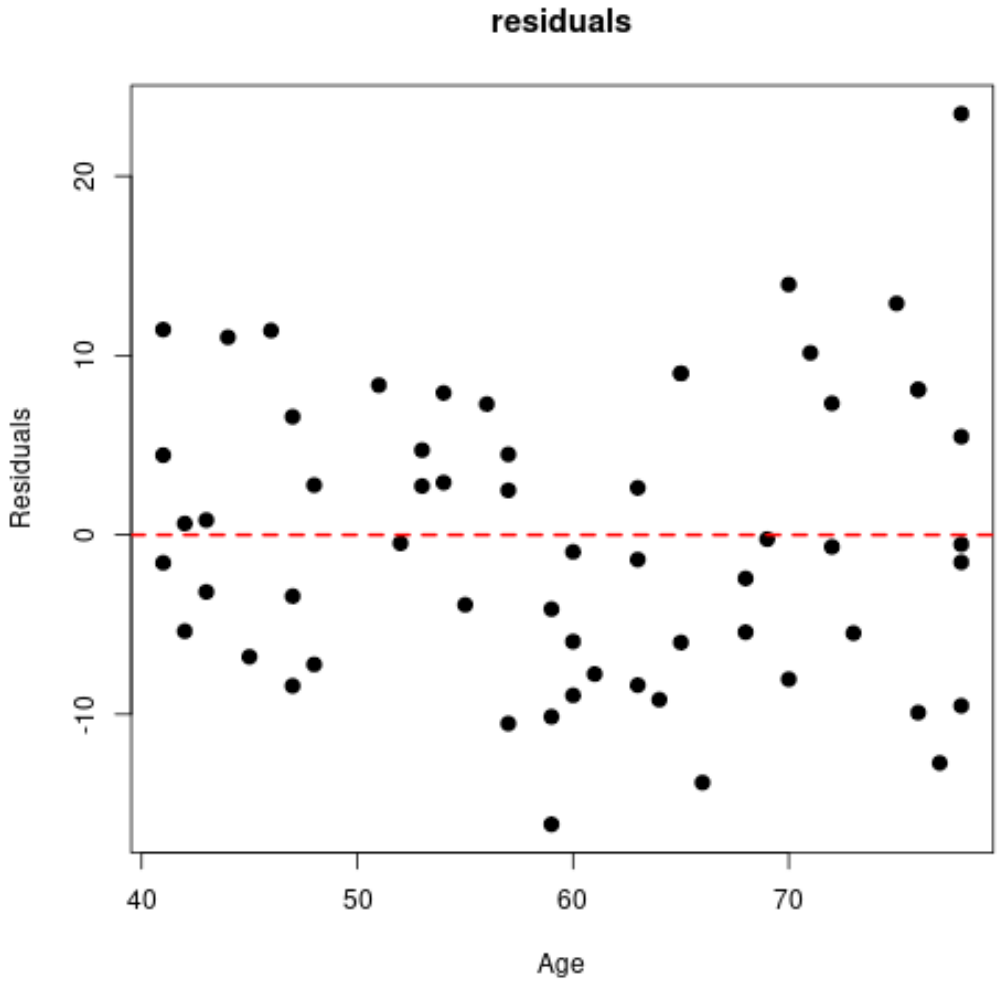
Model Assumptions and Diagnostics

The assumptions in regression are:

- $\epsilon_i \sim N(0, \sigma^2)$
- The values of ϵ are all independent, so that ϵ_i does not depend on ϵ_j
- We assume that the independent variable X is *linearly related* to the dependent variable Y
- The variability of ϵ does not change as X changes; this is known as *homoscedasticity*

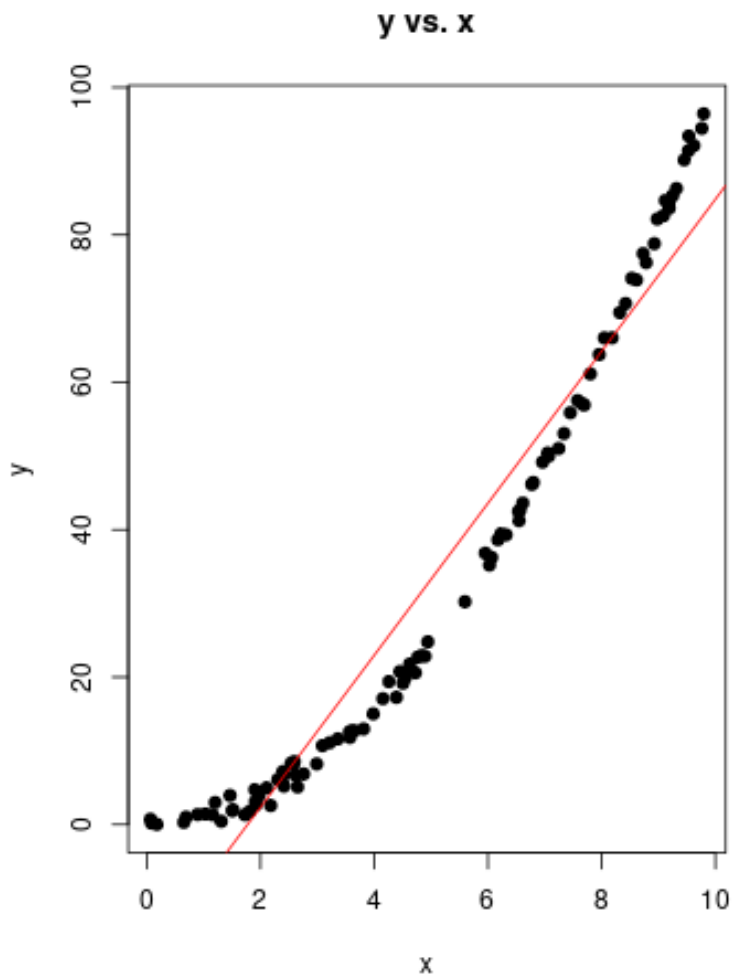
Assumptions - Residuals

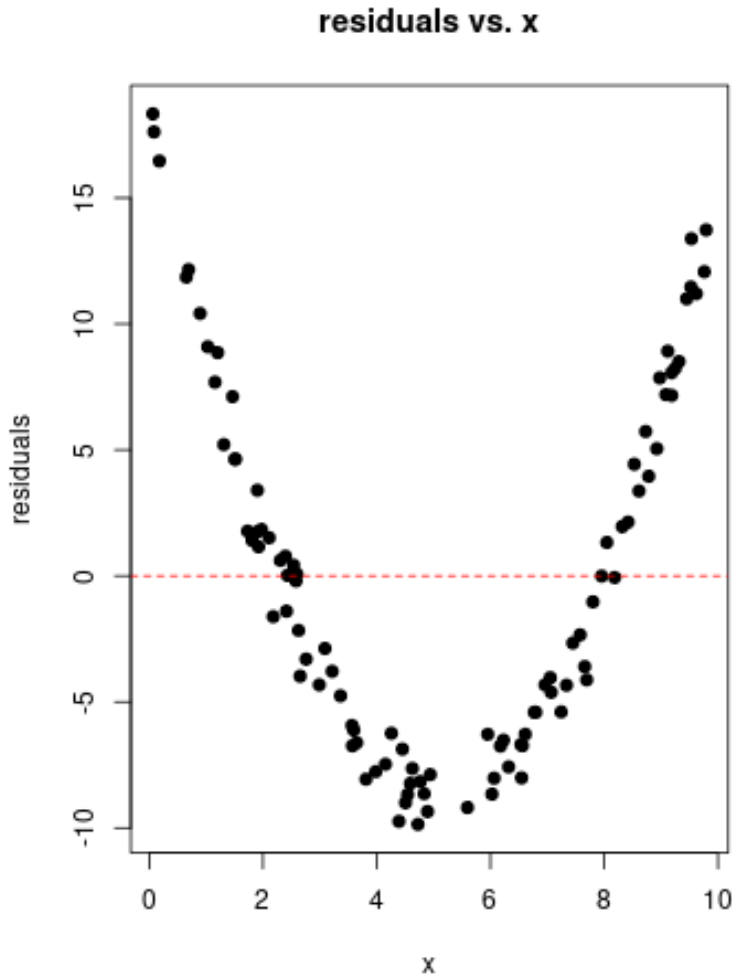
We can examine the first assumption by plotting the residuals from our muscle mass model: we should expect that for every value of x, the residuals are centered around 0



Assumptions - Residuals

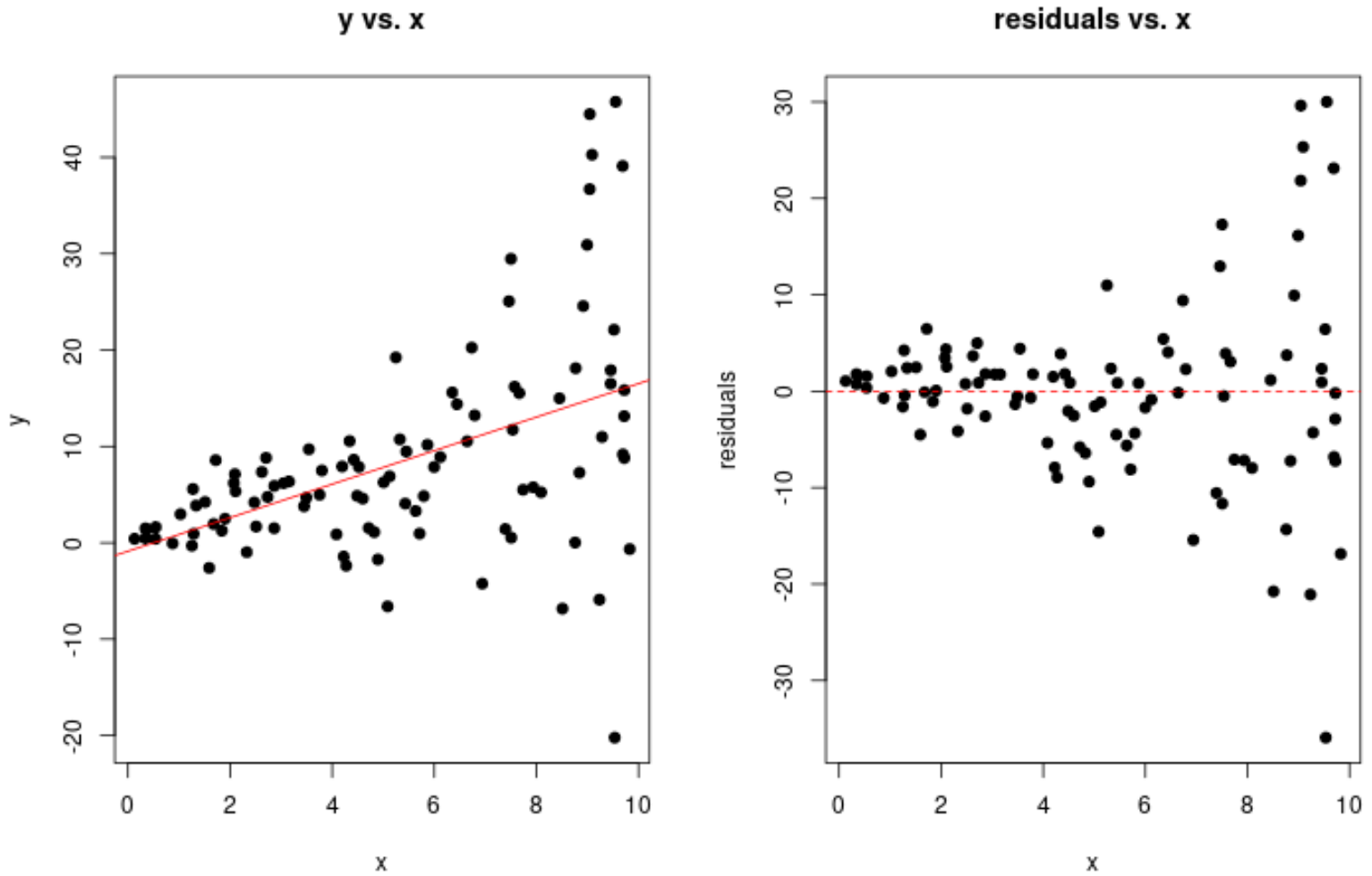
When X and Y do not have a linear relationship, it will often be reflected in the residual plot





Assumptions - Residuals

We will see a similar issue if the value of the residuals does depend on X. This violates the assumption of homoscedasticity:



Remarks on Assumptions

While there are qualtiative tools available to precisely measure these deviances, there are no hard and fast rules for the model assumptions

Whether or not the model remains valid is to some degree subjective – however, being aware of where the assumptions fail may give insight into a problem

For example, if the residuals suggest that there is not a linear relationship, perhaps transformaing the variables or help, or possibly important variables are missing from the model

Review - Model

1. Investigating linear relationship between X and Y: $Y = \beta_0 + \beta_1 X + \epsilon$
2. Choose values of β_0 and β_1 to minimizie $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (residual sum of squares)
3. While this model can be used for prediction, we must stay within range of observed values – intercept value only makes sense if 0 is included in the range of X
4. Estimates of β follow a t distribution
5. $R^2 = 1 - \frac{SS_{Total}}{SS_{Residual}}$ used to determine proportion of variance explained in the model

Review - Residuals

1. Assume that errors are independent of X and each other, with $\epsilon \sim N(0, \sigma^2)$
2. We can often check model assumptions by plotting residuals against X
3. If X and Y do not have a linear relationship, this will be reflected in residual plot. Similarly if variance in ϵ changes with X
4. Model assumptions often don't have hard and fast rules – context and judgement is necessary

References

- Caitlin Ward BIOS:4120 Course Notes