

Session regression I: simple linear regression

Learning outcomes

- understand simple linear regression model incl. terminology and mathematical notations
 - estimate model parameters and their standard error
 - use model for checking the association between x and y
 - use model for prediction
 - assess model accuracy with RSE and R^2
 - check model assumptions
 - to be able to use `lm` function in R for model fitting, obtaining confidence interval and predictions
-

Introduction

Quiz: What do we already know about `simple linear regression`?

Description

- Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative, numerical) variables
 - one variable, denoted x is regarded as the *predictor*, *explanatory*, or *independent variable*, e.g. body weight (kg)
 - the other variable, denoted y , is regarded as the *response*, *outcome*, or *dependent variable*, e.g. plasma volume (liters)
- It is used to estimate the best-fitting straight line to describe the association

Used for to answer questions such as:

- is there a relationship between x exposure (e.g. body weight) and y outcome (e.g. plasma volume)?
- how strong is the relationship between the two variables?
- what will be a predicted value of the y outcome given a new set of exposure values?
- how accurately can we predict the outcome?

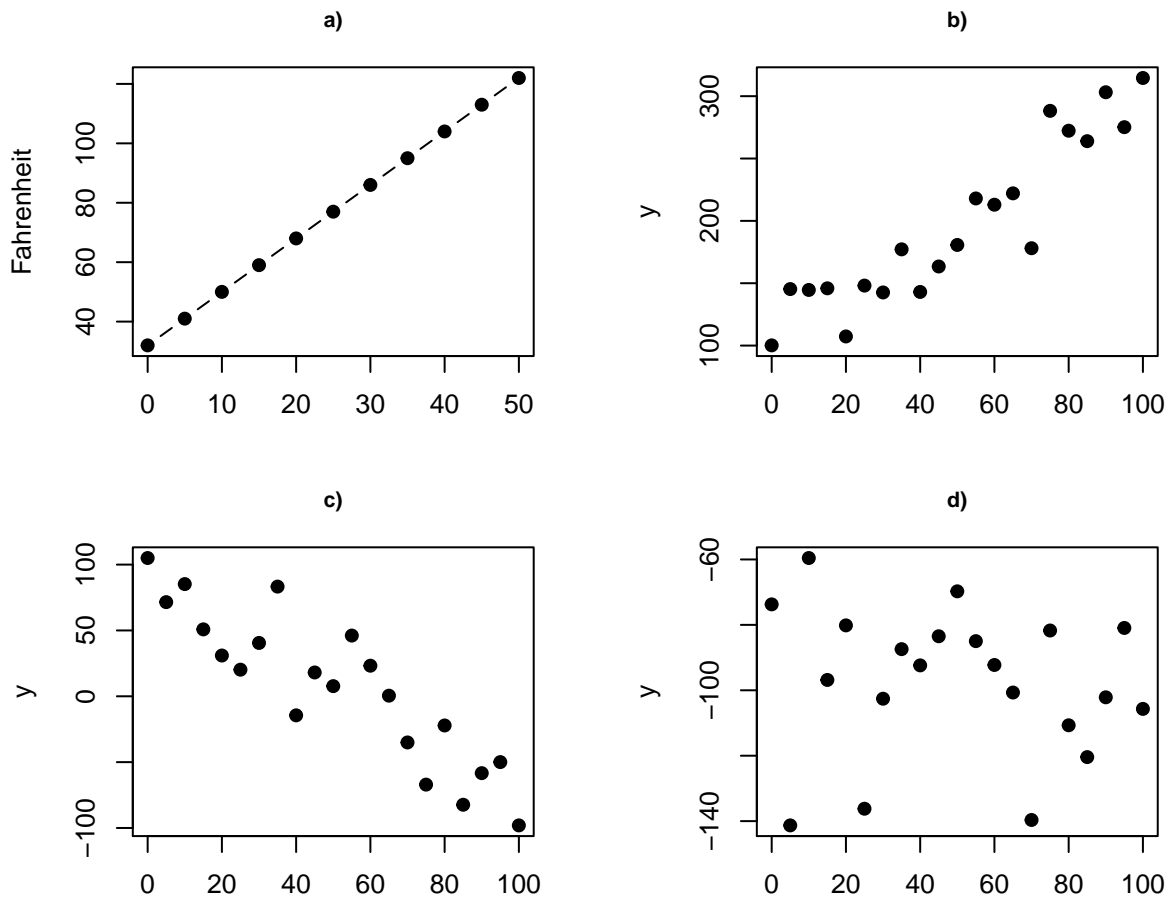


Figure 1: Deterministic vs. statistical relationship: a) deterministic: equation exactly describes the relationship between the two variables e.g. $Fahrenheit = 9/5 * Celsius + 32$; b) statistical relationship between x and y is not perfect (increasing), c) statistical relationship between x and y is not perfect (decreasing), d) random signal

Example data

Example data contain the body weight (kg) and plasma volume (liters) for eight healthy men.

```
weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0) # body weight (kg)
plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12) # plasma volume (liters)
```

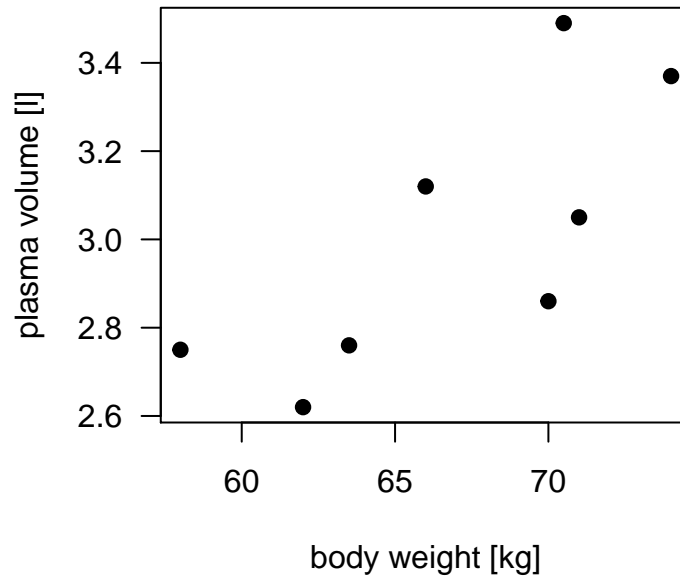


Figure 2: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice versa*.

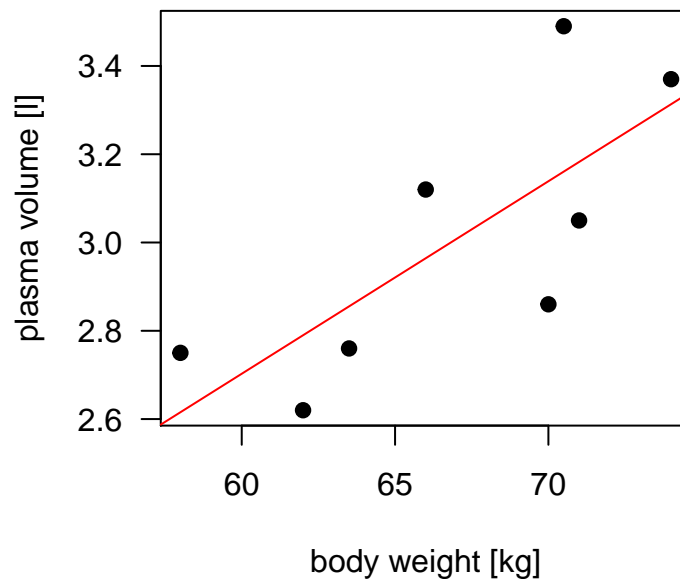


Figure 3: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice versa*. Linear regression gives the equation of the straight line that best describes how the outcome changes (increase or decreases) with a change of exposure variable (in red)

The equation of the regression line is:

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

or mathematically using matrix notation

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

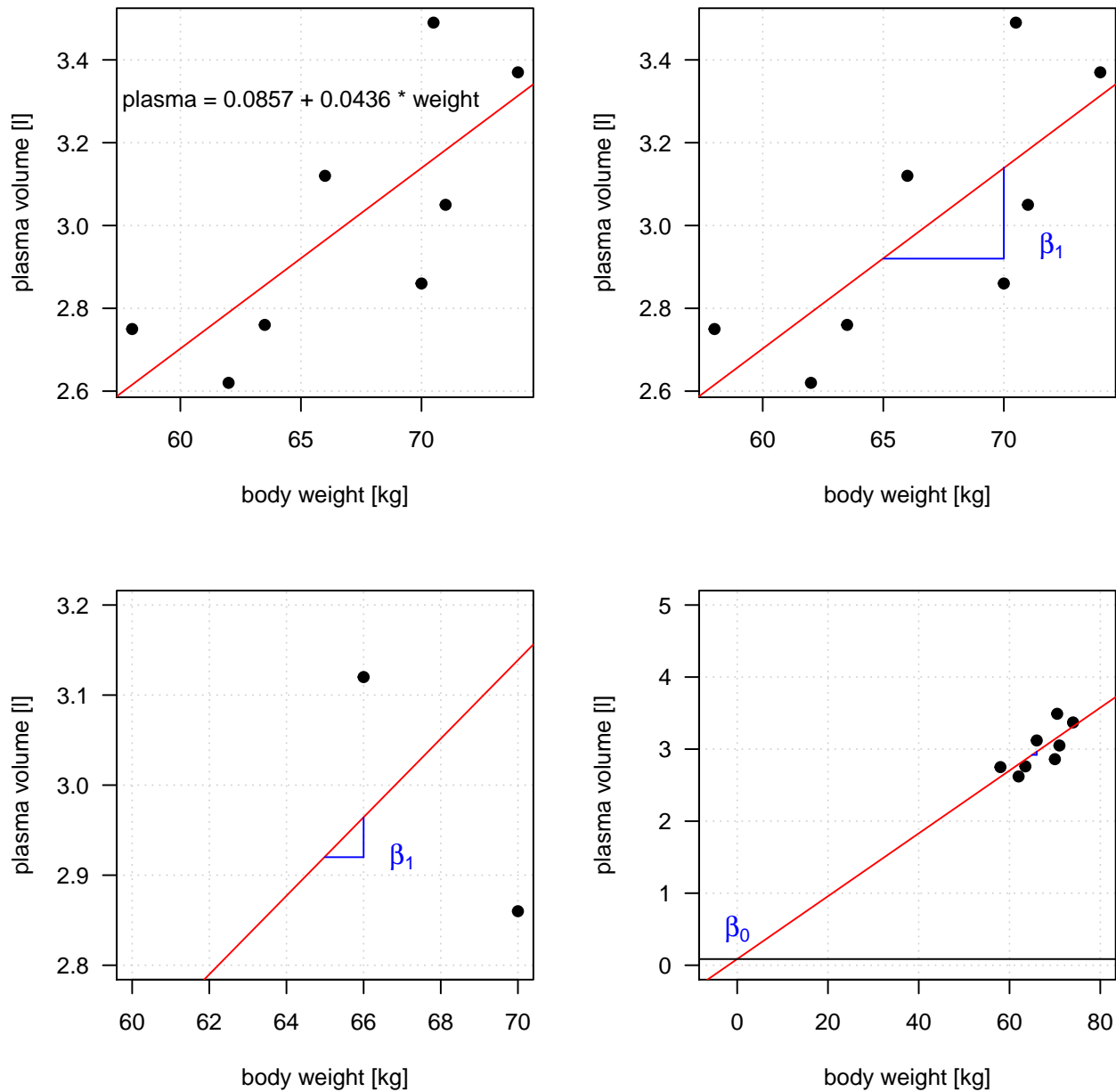
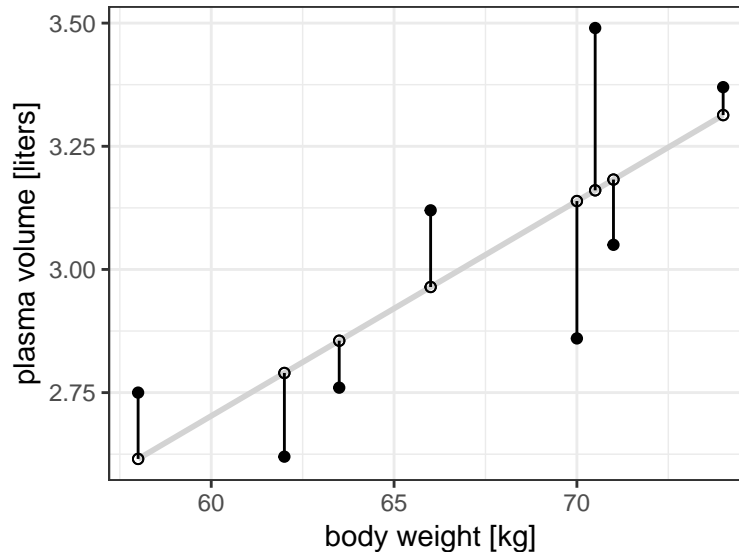


Figure 4: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice versa*. Linear regression gives the equation of the straight line that best describes how the outcome changes (increase or decreases) with a change of exposure variable (in red). Parameters explanation

Quiz: regression model parameters

Estimating model coefficients

In practice, β_0 and β_1 are usually unknown. The best-fitting line is derived using the method of **least squares**, i.e. by finding the values of the parameters β_0 and β_1 that minimize the sum of the squared vertical distances of the points from the line.



Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ represent n observation pairs, each of which consists of a measurement of X and Y , e.g. in our example we have 8 pairs of observations, e.g. $(58, 2.75)$, $(70, 2.86)$ etc.

```
weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0) # body weight (kg)
plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12) # plasma volume (liters)
```

We seek to find coefficients estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that linear model fits the available data well, i.e. such that the resulting line is as close as possible to the 8 data points.

There are a number of ways of measuring *closeness*. By far the most common approach involves minimizing the *least squares* criterion.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction Y based on the i th value of X . Then $\epsilon_i = y_i - \hat{y}_i$ represents the i th *residual*, i.e. the difference between the i th observed response value and the i th response value that is predicted by the linear model.

RSS, the *residual sum of squares* is defined as:

$$RSS = \epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_n^2$$

or equivalently as:

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. With some calculus one gets:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

[Pen and Paper exercise](#): Estimating model coefficients

Hypothesis testing

Accuracy of the coefficient estimates

- The calculated $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of the population values of the intercept and slope and are, therefore, subject to sampling variation
- Their precision is measure by their standard errors

$$s.e(\hat{\beta}_0) = s * \sqrt{\left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}$$

$$s.e(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

where, s is the *standard deviation of the points about the line*. It has $(n - 2)$ degrees of freedom, i.e. the sample size minus the number of regression coefficients

$$s = \sqrt{\left[\frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \bar{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2}{n - 2}\right]}$$

Pen and Paper exercise: Accuracy of the coefficient estimates

Confidence interval

- Standard errors can be used to compute **confidence interval**.
- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- The range is defined in terms of lower and upper limits computed from the data. For linear regression, the 95% confidence intervals takes form:

$$[\hat{\beta}_1 - 2 * s.e.(\hat{\beta}_1), \hat{\beta}_1 + 2 * s.e.(\hat{\beta}_1)]$$

and

$$[\hat{\beta}_0 - 2 * s.e.(\hat{\beta}_0), \hat{\beta}_0 + 2 * s.e.(\hat{\beta}_0)]$$

Hypothesis testing

- Standard errors can also be used to perform **hypothesis testing** on the coefficients.
- The most common hypothesis test involves testing the **null hypothesis** of:

H_0 : There is no relationship between X and Y

versus the **alternative hypothesis**

H_a : There is some relationship between X and Y

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_0 : \beta_1 \neq 0$$

since if

$$\beta_1 = 0$$

then the model

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

reduces to

$$Y = \beta_0 + \epsilon$$

To test the null hypothesis we need to determine whether $\hat{\beta}_1$, our estimate of β_1 , is sufficiently far from zero that we can be confident that β_1 is non-zero.

How far is far enough? This depends on the accuracy of $\hat{\beta}_1$, that is standard error $s.e.(\hat{\beta}_1)$. If $s.e.(\hat{\beta}_1)$ is small, then small values of $\hat{\beta}_1$ may provide strong evidence that $\hat{\beta}_1 \neq 0$ and *vice versa*. In practice, we compute a **t-statistics** given by

$$t = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)}$$

which measures the standard deviations that $\hat{\beta}_1$ is away from 0.

If there really is no relationship between X and Y , then we this will have a t -distribution with $n - 2$ degrees of freedom. From previous sessions, we now know how to compute probability of observing any value equal to $|t|$. We call this probability the $p - value$.

We can interpret the $p - value$ as follows: a small p-value indicates that it is unlikely to observe such a substantial association between X and Y due to chance, i.e. in the absence of any real association. We therefore can **reject the null hypothesis**.

Typical p-value cutoffs for rejecting the null hypothesis are 5 or 1%.

[Pen and Paper exercise](#): Hypothesis testing

Live coding demo

```
# Data
weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0)
plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12)

# Plot
plot(weight, plasma)

# Regression
reg <- lm(plasma~weight)
summary(reg)

# Coefficients
coef(reg)

# Confidence intervals
confint(reg)

# Add regression line to the plot
abline(reg)
```

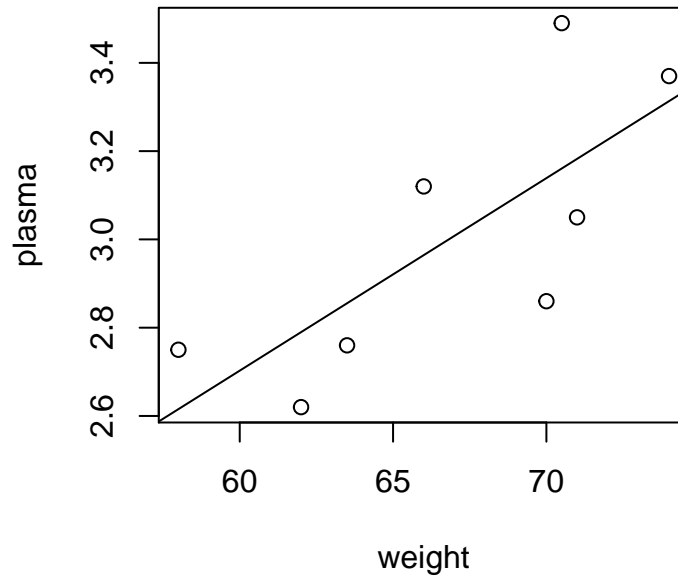


Figure 5: Body weight vs. plasma volume

```
##
## Call:
## lm(formula = plasma ~ weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27880 -0.14178 -0.01928  0.13986  0.32939
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08572    1.02400   0.084   0.9360
## weight       0.04362    0.01527   2.857   0.0289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2188 on 6 degrees of freedom
## Multiple R-squared:  0.5763, Adjusted R-squared:  0.5057
## F-statistic:  8.16 on 1 and 6 DF,  p-value: 0.02893
##
## (Intercept)      weight
##  0.08572428  0.04361534
##              2.5 %    97.5 %
## (Intercept) -2.419908594 2.59135716
## weight      0.006255005 0.08097567
```

Prediction

Sometimes it may be useful to use the regression equation to predict the value of y_i for a particular value of x_i , say x_i^t . The predicted value is:

$$y'_i = \hat{\beta}_0 + \hat{\beta}_1 x'_i$$

and its standard error is:

$$s.e.(y'_i) = s \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

The standard error is least when x_i is close to the mean, \bar{x}

In general, one should be reluctant to use the regression line for predicting values outside the range of x in the original data, as the linear relationship will not necessarily hold true beyond the range over which it has been fitted.

Prediction interval

There is also a concept called **prediction interval**. Here, we look at any specific value of x_i , and find an interval around the predicted value y'_i for x_i such that there is a 95% probability that the real value of y (in the population) corresponding to x_i is within this interval.

Prediction interval regression vs. confidence interval

- 95% confidence interval: there is 95% probability that the true best fit-line for the population lies within the confidence interval
- 95% prediction interval: 95% of the y values found for a certain x value will be within the interval range around the linear regression line
- prediction interval > than a confidence interval, as it must account for both the uncertainty in knowing the value of the population mean, plus data scatter.

The 95% prediction interval of the forecasted value y'_i :

Live coding demo

```
# Prediction
predict(reg, data.frame(weight=60))

##          1
## 2.702645

predict(reg, data.frame(weight=c(60, 70)))

##          1          2
## 2.702645 3.138798

# Prediction with confidence intervals
predict(reg, data.frame(weight=66), interval="prediction")

##          fit          lwr          upr
## 1 2.964337 2.395511 3.533162
```

Assessing the Accuracy of the Model & Correlation

Once we have rejected the null hypothesis (H_0 : there is no relationship between X and Y) in favor of the alternative hypothesis (H_a : there is some relationship between X and Y) we may want to quantify **the extent to which the model fits the data**.

The quality of a linear regression fit is typically assessed using two related quantities: - RSE, the residual standard error - R^2 statistics

RSE, Residual standard error

- RSE is a measure of **lack of fit** of the model to the data
- It is measured in units of Y .

Going back to linear regression model, and writing it in the formal complete way:

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

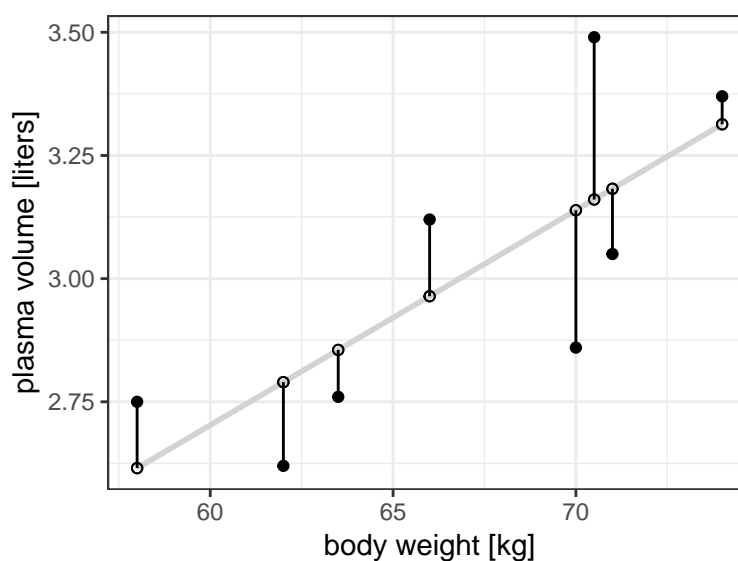
we can see that each observation is associated with an error term ϵ . This means that even knowing β_0 and β_1 one cannot perfectly predict Y from X .

RSE is an estimate of the standard deviation of ϵ , that can be viewed as the average amount that the response will deviate from the true regression line. It is calculated

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

where

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



Live coding demo

```
#weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0) # body weight (kg)
#plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12) # plasma volume (liters)
```

```
plasma
```

```
## [1] 2.75 2.86 3.37 2.76 2.62 3.49 3.05 3.12
```

```
weight
```

```
## [1] 58.0 70.0 74.0 63.5 62.0 70.5 71.0 66.0
```

```
head(data.reg)
```

```
##   plasma weight predicted   residuals
## 1   2.75   58.0  2.615414  0.13458612
## 2   2.86   70.0  3.138798 -0.27879793
## 3   3.37   74.0  3.313259  0.05674072
## 4   2.76   63.5  2.855298 -0.09529823
## 5   2.62   62.0  2.789875 -0.16987523
## 6   3.49   70.5  3.160606  0.32939440
```

```
reg <- lm(data.reg$plasma~data.reg$weight)
```

```
# predict Y given the values of X and regression model reg
```

```
y.pred <- predict(reg, data.frame(weight=data.reg$weight))
y.pred
```

```
##           1           2           3           4           5           6           7           8
## 2.615414  3.138798  3.313259  2.855298  2.789875  3.160606  3.182413  2.964337
```

```
# calculate residuals
```

```
e.terms <- data.reg$plasma-y.pred
e.terms
```

```
##           1           2           3           4           5           6
## 0.13458612 -0.27879793  0.05674072 -0.09529823 -0.16987523  0.32939440
##           7           8
## -0.13241327  0.15566342
```

```
# calculate RSS
```

```
RSS=sum(e.terms^2)
```

```
# calculate RSE
```

```
n=nrow(data.reg)
```

```
RSE <- sqrt((1/(n-2))*RSS)
```

```
# R reg objects contains it all
```

```
names(reg)
```

```
## [1] "coefficients" "residuals"      "effects"        "rank"
## [5] "fitted.values" "assign"         "qr"            "df.residual"
## [9] "xlevels"      "call"          "terms"         "model"
```

```
reg$fitted.values
```

```
##           1           2           3           4           5           6           7           8
## 2.615414  3.138798  3.313259  2.855298  2.789875  3.160606  3.182413  2.964337
```

```
reg$residuals
```

```
##           1           2           3           4           5           6
## 0.13458612 -0.27879793  0.05674072 -0.09529823 -0.16987523  0.32939440
##           7           8
## -0.13241327  0.15566342
```

```
# RSE
```

```
summary(reg)
```

```
##
```

```
## Call:
## lm(formula = data.reg$plasma ~ data.reg$weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27880 -0.14178 -0.01928  0.13986  0.32939
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.08572     1.02400   0.084   0.9360
## data.reg$weight  0.04362     0.01527   2.857   0.0289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2188 on 6 degrees of freedom
## Multiple R-squared:  0.5763, Adjusted R-squared:  0.5057
## F-statistic:  8.16 on 1 and 6 DF,  p-value: 0.02893
```

R^2 statistics

- R^2 statistics is an alternative measure of fit and measure of linear relationship between X and Y
- It takes the form of a proportion, the proportion of variance explained, hence is independent of the scale of Y
- $0 \leq R^2 \leq 1$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Correlation

- is also a measure of linear regression between X and Y

$$Cor(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

This suggests that we should be able to use $r = Cor(X, Y)$ to assess the fit of the linear model

In fact, For simple linear regression, it can be shown that

$$R^2 = r^2$$

Assumptions