Simple Linear Regression Assignment

Maxwell Maia, id=21236277

26 November, 2022

Predicting 3 km Running Times based on laboratory testing.

Study Description:

Sixteen well-trained male middle and long distance runners performed a 3 km time trial and a number of running tests in the laboratory including their running velocity at a blood lactate concentration of 4 mmol.l-1 (v4mM). Other variables measured were running velocity at their Lactate Threshold (vTlac), and VO2 max. All the laboratory testing took place on a motorised treadmill, and distance running performance was determined by 3 km time trials on an indoor 200m track.

Aims:

To investigate whether there is we can use linear regression to predict 3 km running time (minutes) from v-4mM (km per hour) in the population of well-trained male middle and long distance runners. Hence to predict 3km running time using running velocity at blood lactate concentration 4 mmol per litre.

- Response Variable: 3km running time (`Running.Time`) measured in minutes
- Explanatory Variable: running velocity at blood lactate concentration at 4mmol per litre (`v4mM`) measured in km/hr

```
library(tidyverse)
```

Read the data and see a few rows

```
running = read.csv("3krunning.csv", header = TRUE)
head(running)
```

```
##   Running.Time v4mM vTlac Rel.14.5 Rel.16.1 VO2Max
## 1         8.23 20.4  19.5     47.1     52.4   23.4
## 2         8.30 19.5  18.2     48.1     60.0   23.5
## 3         8.62 19.0  17.3     50.3     56.8   22.0
## 4         8.82 18.9  17.8     51.8     56.1   23.0
## 5         9.18 17.8  16.5     48.7     54.1   21.5
## 6         9.23 17.2  15.6     50.5     59.6   20.5
```

Summary Statistics

Task: Calculate the summary statistics for each column in the data and describe the key features of the data.

```
running %>%
  summarize(Mean.RunningTime = mean(Running.Time),
            SD.RunningTime = sd(Running.Time),
            Mean.v4mM = mean(v4mM),
            SD.v4mM = sd(v4mM),
            Mean.vTlac = mean(vTlac),
            SD.vTlac = sd(vTlac),
            Mean.Rel.14.5 = mean(Rel.14.5),
            SD.Rel.14.5 = sd(Rel.14.5),
            Mean.Rel.16.1 = mean(Rel.16.1),
            SD.Rel.16.1 = sd(Rel.16.1),
            Mean.VO2Max = mean(VO2Max),
            SD.VO2Max = sd(VO2Max))
```

```
##   Mean.RunningTime SD.RunningTime Mean.v4mM  SD.v4mM Mean.vTlac SD.vTlac
## 1         9.458125       0.744269  17.06875 1.848141      15.95 1.775763
##   Mean.Rel.14.5 SD.Rel.14.5 Mean.Rel.16.1 SD.Rel.16.1 Mean.VO2Max SD.VO2Max
## 1      51.59375    3.289877      57.81875    3.775221     20.6875  2.133503
```

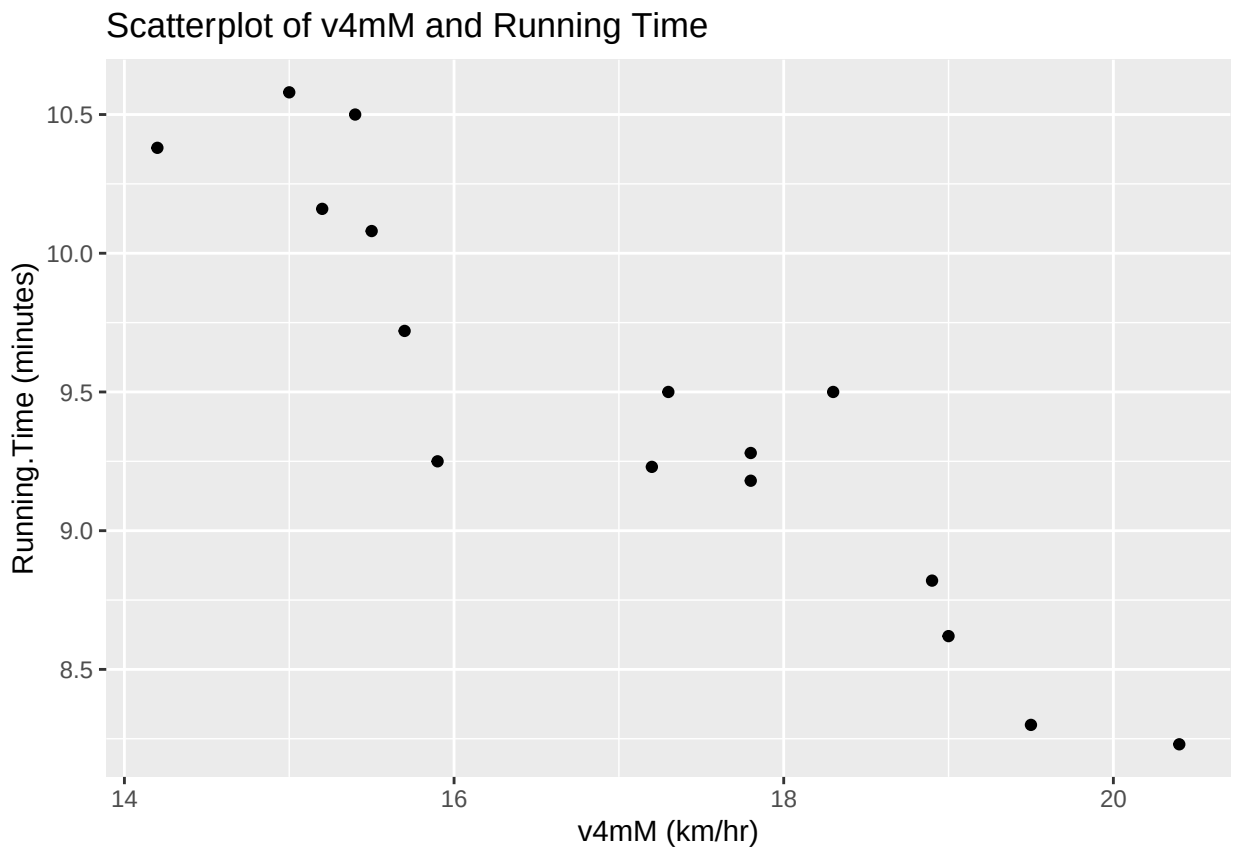
The mean of the variables shows is a single value indicator of the statistics of a runner for all of these fields. They will be used to calculate the standard deviation. The mean of `Running.Time` is 9.458125. The mean of `v4mM` is 17.06875.

The running time has a relatively low standard deviation of 0.74 which means that the running times of the 16 male runners are clustered around the mean. This may be because all of the runners are well-trained so the difference in performance is minimal. The standard deviation of the v4mM variable is higher than that of the running time variable. The v4mM standard devation is 1.85. The data of the v4mM variable is more spread out than the running variable, but not too much more spread out. This increase in spread can be attributed to people having different balances of chemicals in their bodies. These athletes don't all eat and live in exactly the same way and their genetics are different too so a higher spread in v4mM is to be expected.

Scatterplot

Task: Make a labelled scatterplot of `v4mM` vs `Running.Time` and interpret it.

```
ggplot(running, aes(y = Running.Time, x = v4mM)) +
  geom_point() +
  labs(x = "v4mM (km/hr)", y = "Running.Time (minutes)",
       title = "Scatterplot of v4mM and Running Time")
```



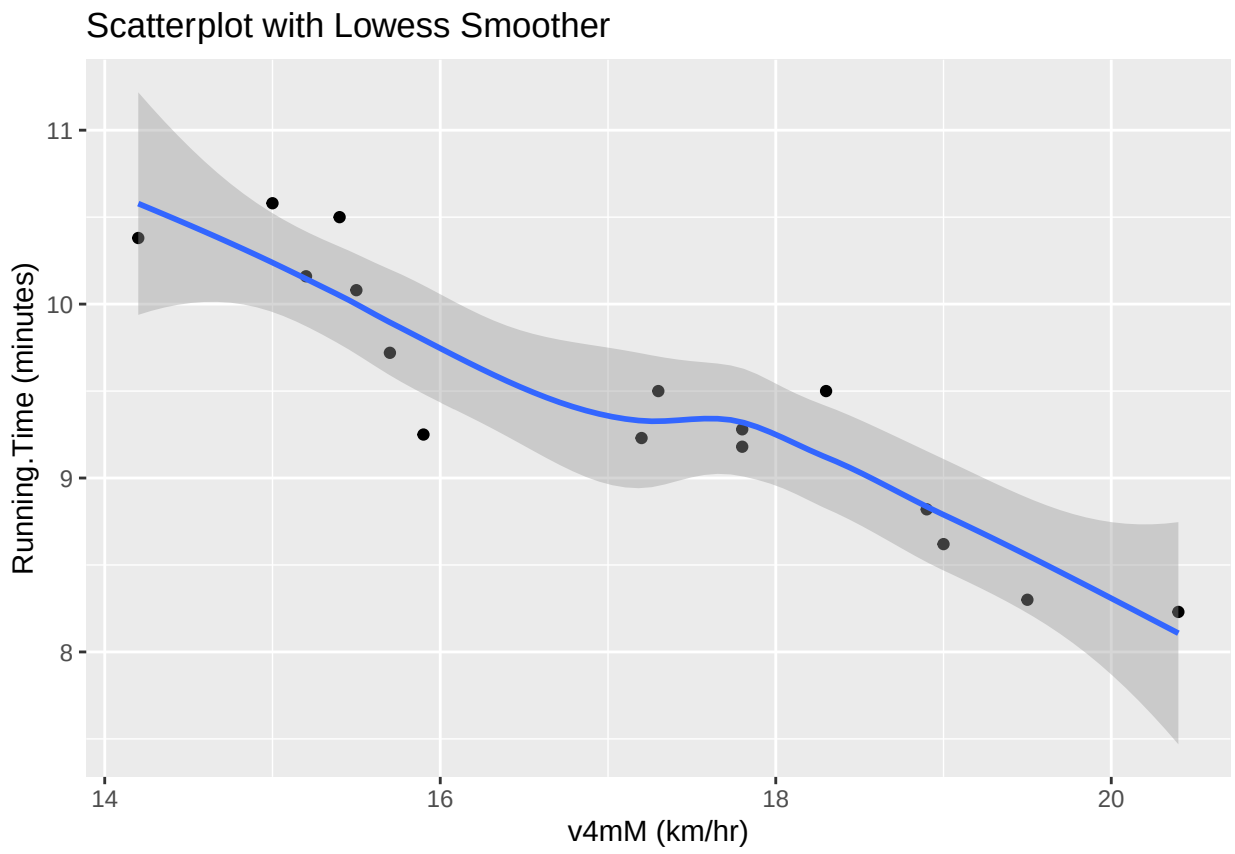
Interpretation: There is a trend that as the v4mM increases, the running time decreases. The data appears to create a linear downward slope. This means that the data seems to have a linear negative association. As the velocity increases, the running time decreases. This makes sense because the faster an athlete runs, the quicker they can complete a race.

Scatterplot with smoother.

Task: Add a smooth line to the scatter plot produced in the previous task, and include the new plot below.

```
ggplot(running, aes(y = Running.Time, x = v4mM)) +
 geom_point() +
 geom_smooth() +
 labs(x = "v4mM (km/hr)", y = "Running.Time (minutes)",
 title = "Scatterplot with Lowess Smoother")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Task: What does the smoother suggest regarding the suitability of a simple linear regression model for this relationship?

The blue line is relatively straight. If you tried to fit a straight line within the grey area, you probably could. That means that is is probably a linear association.

Correlation coefficient

Task: calculate the correlation coefficient between `v4mM` vs `Running.Time` and interpret it.

```
running %>% select (Running.Time, v4mM) %>% cor()
```

```
##              Running.Time      v4mM
## Running.Time     1.000000 -0.925857
## v4mM            -0.925857  1.000000
```

The correlation coefficient gives an indiction of the strength of the linear relationship between 2 variables. r = -0.925857. There is a strong, negative correlation coefficient. This means that there is a strong, negative linear relationship between v4mM and Running time. As v4mM increases, running time decreases. This is assuming that:

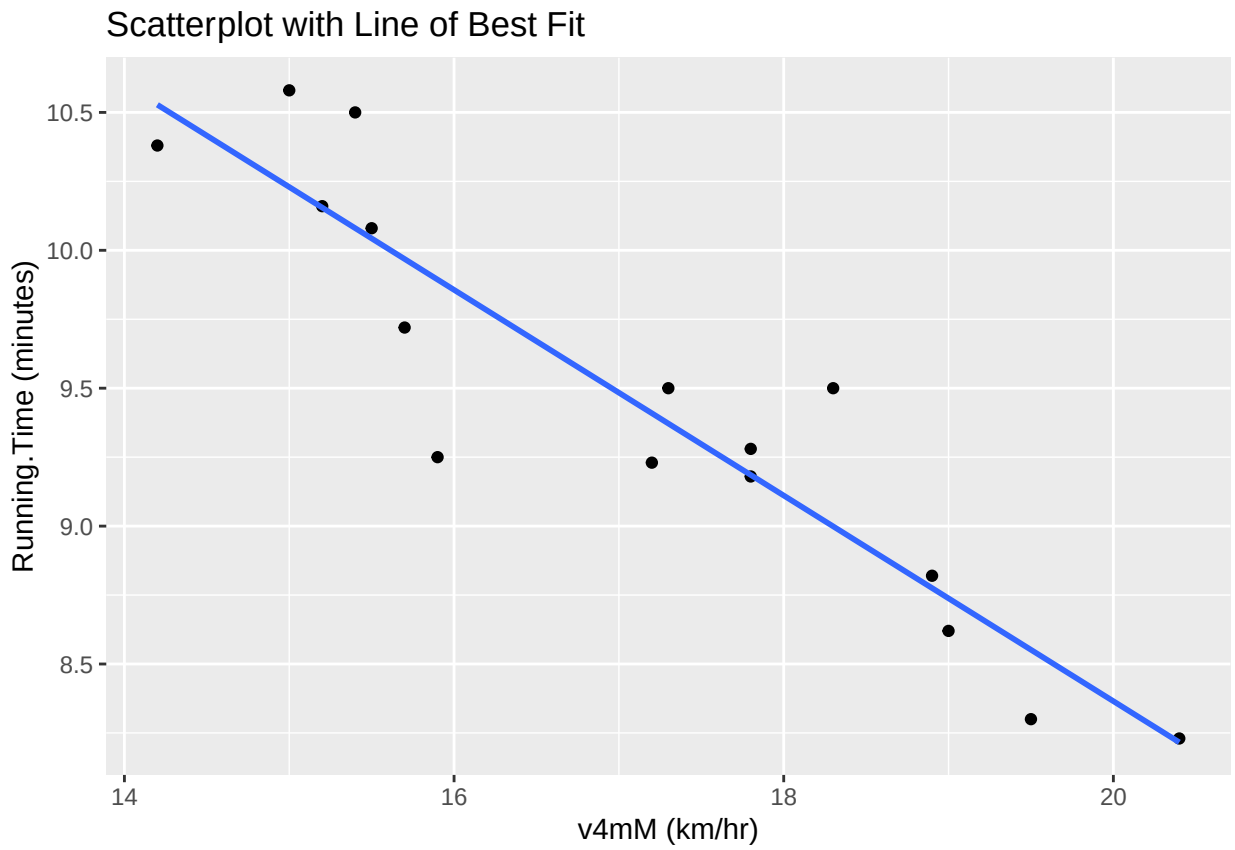
There is linearity; by visual assessment: there is linearity. Independent observations. Normal residuals. Points are evenly scattered around the line.

Scatterplot with line of best fit

Task: Add the line of best fit to the scatter plot produced above and interpret it.

```
ggplot(running, aes(y = Running.Time, x = v4mM)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)+
  labs(x = "v4mM (km/hr)", y = "Running.Time (minutes)",
       title = "Scatterplot with Line of Best Fit")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The line of best fit attempts to pass as close through all of the data as possible. There are no groupings in this data so a line of best fit that goes from the top left of the graph to the bottom right of the graph is appropriate. The slope of the line of best is negative which aligns with the negative of the correlation coefficient.

Fitting a Simple Linear Regression Model

Task: Estimate the parameters of the line of best fit for the linear relationship between 3 km running time and v-4mM in the sample. This can then be used for inference about the linear relationship in the population of well-trained male middle and long distance runners.

```
running.model = lm(Running.Time ~ v4mM, running)
running.model
```

```
##
## Call:
## lm(formula = Running.Time ~ v4mM, data = running)
##
## Coefficients:
## (Intercept)         v4mM
##     15.8223      -0.3729
```

Equation of line of best fit

Task: Write down the equation of the line of best fit and also provide an interpretation of the slope and intercept. Does the intercept have a physically meaningful interpretation?

```
#Find the slope and intercept.

#using R
lm(Running.Time ~ v4mM, running)
```

```
##
## Call:
## lm(formula = Running.Time ~ v4mM, data = running)
##
## Coefficients:
## (Intercept)         v4mM
##     15.8223      -0.3729
```

```
#How to calculate by hand.
#Slope = r * Sy/Sx
#slope = ....

#Then
#mean(y) = intercept + slope*mean(x)
#Sub in mean(y), slope and mean(x)
```

y = 15.8223 - 0.3729x

The slope is the measure of how rapidly the Predicted value of y ($\hat{y}$) changes with respect to x. In this study the slope is the change of running time per unit increase in v4mM. How much quicker in the 3km time trial is the data subject per unit increase of their velocity at a blood lactate concentration of 4 mmol/1. If the runner has 1 higher km/hr in their v4mM variable then they will have a decrease of 0.3729 minutes in their 3km running time.

The intercept would be the time that a runner takes to complete a 3km marathon of a runner that moves at 0 velocity when their blood lactate concentration is 4mmol/1. He probably wouldn't be healthy enough to run, which isn't relevant to this study. The intercept does not have a physically meaningful interpretation.

Make some point predictions

Task: Predict the running time (i.e. `Running.Time`) when running speed at blood lactate concentration 4 mmol/litre (i.e. `v4mM`) are 14, 15, 16, 17, 18, 19 and 20 km per hour.

```r
# New data points for explanatory variable that we will get a prediction for.
running_new = data.frame(
  v4mM = c(14, 15, 16, 17, 18, 19, 20)
)
```

Numbered 1 - 7 in the are the predicted 3km running times when v4mM is 14, 15, 16, 17, 18, 19 and 20 respectively. This is based of the the line of best fit.

```r
predict(running.model, newdata = running_new, interval = "none")
```

```
##         1         2         3         4         5         6         7
## 10.602320 10.229467  9.856613  9.483759  9.110905  8.738051  8.365197
```

Interval estimation for predicted running times

For each of the predictions produce a 95% confidence interval and 95% prediction interval, and interpret the results carefully.

95% confidence interval: a range of values that is likely to contain the true mean value of the response variable, given a specific value of the explanatory variable. This does not tell you about the spread of the individual data points around the true mean.

Numbered 1 to 7, are the 95% confidence intervals (lwr, upr) for each prediction.

```r
predict(running.model, newdata = running_new, interval = "confidence", level = 0.95)
```

```
##         fit       lwr       upr
## 1 10.602320 10.292450 10.912191
## 2 10.229467  9.990868 10.468065
## 3  9.856613  9.674799 10.038426
## 4  9.483759  9.327551  9.639966
## 5  9.110905  8.934940  9.286869
## 6  8.738051  8.508390  8.967712
## 7  8.365197  8.065626  8.664767
```

95% Prediction interval: a range of values that is likely to contain the actual value of the response variable (for a single new observation, given a specific value of the explanatory variable.). The prediction interval is for individual oberservations rather than the mean.

Numbered 1 to 7, are the 95% prediction intervals (lwr, upr) for each prediction.

```r
predict(running.model, newdata = running_new, interval = "prediction", level = 0.95)
```

```
##         fit      lwr       upr
## 1 10.602320 9.905285 11.299356
## 2 10.229467 9.561059 10.897874
## 3  9.856613 9.206309 10.506916
## 4  9.483759 8.840144 10.127373
## 5  9.110905 8.462212  9.759598
## 6  8.738051 8.072781  9.403320
## 7  8.365197 7.672678  9.057715
```

Plots with confidence and prediction intervals

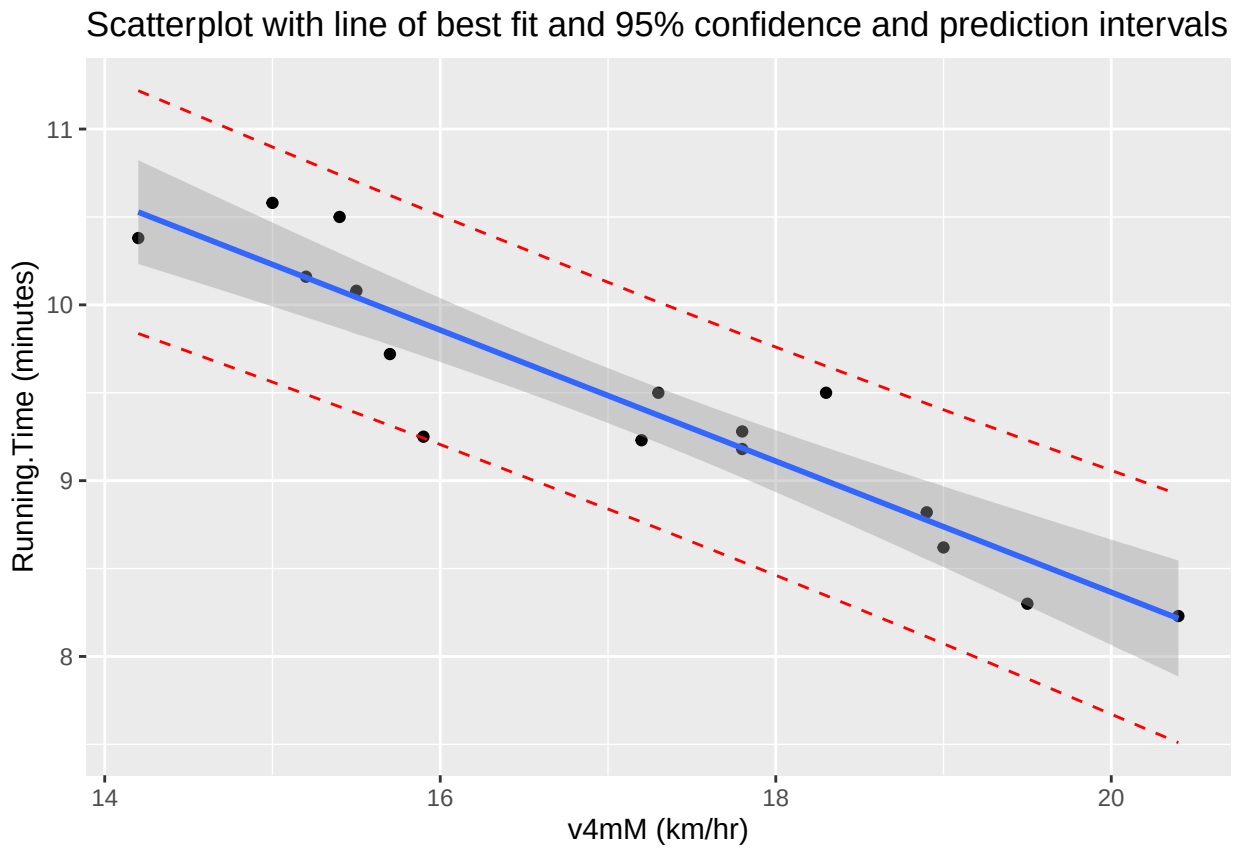
Task: Add the 95% confidence and 95% prediction intervals to the scatter plot with the line of best fit, and interpret.

```
pred.int = predict(running.model, newdata = running, interval = "prediction")

running2 = cbind(running, pred.int) # store predictions alongside original dataset

ggplot(running2, aes(y = Running.Time, x = v4mM)) +
  geom_point() +
  stat_smooth(method = lm) + # this includes a confidence interval
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") + #  Add prediction intervals
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  labs(x = "v4mM (km/hr)", y = "Running.Time (minutes)",
       title = "Scatterplot with line of best fit and 95% confidence and prediction intervals")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Interpretation:

More prediction

Task: Predict the running time (i.e. `Running.Time`) when running speed at blood lactate concentration 4 mmol/litre (i.e. `v4mM`) is 18.9 km per hour.

```r
more.prediction.value <- data.frame(v4mM = 18.9)

predict(running.model, newdata = more.prediction.value, interval = "none")
```

```
##        1
## 8.775336
```

Task: Why is the result here is different from 8.82, the observed running time when running speed at blood lactate concentration 4 mmol/litre (`v4mM`) is 18.9 mmol.l-1? (see observation row 4)

The result of the prediction was different to that of which was observed because this prediction was made from the line of best fit that was calculated with this sample set. The line of best fit is the line for which the sum of the squared residuals is smallest. This means that the line of best fit does not go exactly over every data point, which means a prediction based off of the same sample data won't always be exactly accurate. In other words, the discrepancy is due to the error term.

Task: Predict the running time (i.e. `Running.Time`) when `v4mM` is 2.6 km per hour. Explain if you have any concern related to this prediction.

```r
more.prediction.value2 <- data.frame(v4mM = 2.6)

predict(running.model, newdata = more.prediction.value2, interval = "none")
```

```
##        1
## 14.85286
```

This prediction is likely inaccurate because we have not observed data when the v4mM is nearly as low as 2.6 km/hr. 2.6 is an outlier to the data that we have considered. We cannot be sure that the association is still linear outside of the range of the explanatory variable that we have determined to be linear. Therefore, it would not be safe (in terms of correctness) to predict a response variable for an explanatory variable of 2.6 using a linear prediction model.

Overall Conclusion

Task: State your overall conclusions from fitting a linear model for the relationship between 3k running time and the running speed at blood lactate concentration 4 mmol/litre.

There is substantial evidence to say that there is an association between 3km running time and the Running velocity at a blood lactate concentration of 4 mmol/1. There is a strong negative correlation between these 2 variables. The association can be linearly modelled. An athlete with a lower v4mM will have a higher running time. If an athlete wanted to reduce their running time they should look into increasing their running speed at a blood lactate concentration of 4 mmol/litre.

We can safely use linear regression to predict the value of 3km running times (minutes) from v-4mM (km per hour) in the population of well-trained male middle and long distance runners. (As long as the explanatory variable used lies within a safe range to make a prediction).