Simple Linear Regression Assignment

eidhne kennedy, id=21347021

28 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.Time = mean(Running.Time),  
 SD.Time = 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.Time SD.Time Mean.v4mM SD.v4mM Mean.vTlac SD.vTlac Mean.Rel.14.5  
## 1 9.458125 0.744269 17.06875 1.848141 15.95 1.775763 51.59375  
## SD.Rel.14.5 Mean.Rel.16.1 SD.Rel.16.1 Mean.VO2Max SD.VO2Max  
## 1 3.289877 57.81875 3.775221 20.6875 2.133503

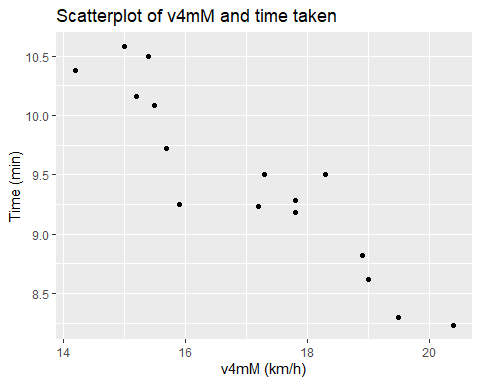
The standard deviation of running time is less than a minute, which should be noted when trying to make any predictions about the running times, as predictions will seem more accurate than they are due to the relative closeness of the sample times. The velocity at blood lactate concentration 4 mmol/litre, however, has a much broader relative standard deviation, which should also be noted.

Another point of note is that these runners are trained for middle- and long-distance running, so there may be a large discrepancy in values like VO2 max, given the specialty of the individual runner and the differences in fitness required for each discipline.

### Scatterplot

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

ggplot(running, aes(x = v4mM, y = Running.Time)) +   
 geom\_point() +  
 labs(x = "v4mM (km/h)", y = "Time (min)",   
 title = "Scatterplot of v4mM and time taken")



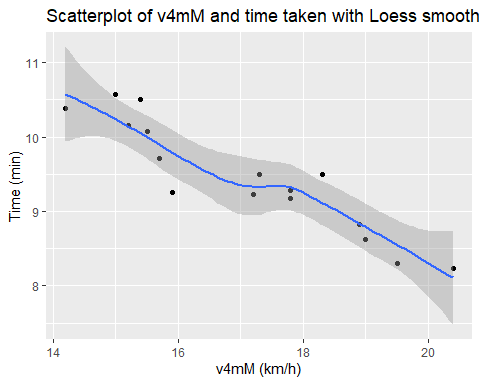
There seems to be a linear, inverse correlation between v4mM and time taken, and there seem to be a few outliers, but it is unclear by eye.

### 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(x = v4mM, y = Running.Time)) +   
 geom\_point() +  
 geom\_smooth() +  
 labs(x = "v4mM (km/h)", y = "Time (min)",   
 title = "Scatterplot of v4mM and time taken with Loess smoother")

## `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 smoother appears to agree with my previous assessment, that there is a clear linear, inverse relationship between these two variables.

It should be noted, however, that there are four clear outliers made visible by the smoother, however despite this the Loess smoother indicates that a linear regression model would be suitable for this data.

### Correlation coefficient

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

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

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

The correlation is very close to -1, therefore there exists a strong inverse correlation between these two variables, as evidenced in the prior scatterplots.

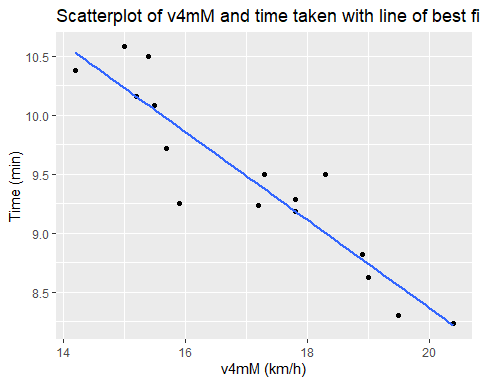
It should be noted, in the face of a correlation this strong, that correlation does not imply causation.

### 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(x = v4mM, y = Running.Time)) +   
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(x = "v4mM (km/h)", y = "Time (min)",   
 title = "Scatterplot of v4mM and time taken with line of best fit")

## `geom\_smooth()` using formula 'y ~ x'



Once again, the line of best fit is clearly demonstrating an inverse linear relationship, with a slope estimated to be around -0.37 units. In addition, most of the points fall very close to the line of best fit, with few outliers, so it seems to be an accurate representation of the overall ttrend of the data.

### 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?

slope = -0.3729

y-intercept = 15.8223

equation = y = -0.3729x + 15.8223

The slope here is significant, as it implies that for every additional km an hour at blood lactate concentration 4mmol/litre that an individual runner is capable of, the runner improves their 3km running time by .37 minutes, which is significant.

The y-intercept does not seem to have a meaningful interpretation, as the time in minutes it would take to run three kilometers when not moving is a nonsense statement.

### 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.

running\_new = data.frame(  
 v4mM = c(14, 15, 16, 17, 18, 18, 20)  
)

### 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.

predict(Running.model, newdata = running\_new, interval = "confidence")

## 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 9.110905 8.934940 9.286869  
## 7 8.365197 8.065626 8.664767

It is likely (p>0.95) that the true mean of recorded times taken to run 3km will lie within the given ranges for each velocity given, however this says nothing about the spread of individual data points that will contribute to this mean.

predict(Running.model, newdata = running\_new, interval = "prediction")

## 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 9.110905 8.462212 9.759598  
## 7 8.365197 7.672678 9.057715

These ranges of values are likely to contain the value of any specific response for a single observation, which is why thr range between the upper and lower bounds are the same.

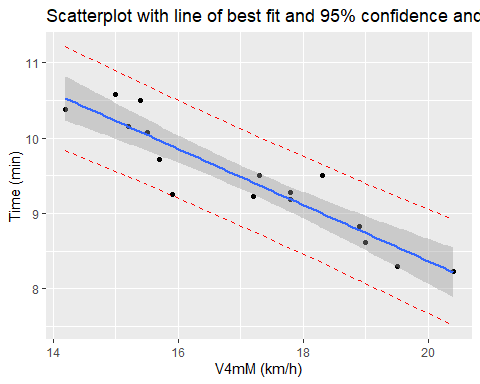
It is worth noting that even the smaller ranges, from the confidence interval, are relatively large (.6 minutes, which is roughly equivalent to 40 seconds), and for athletes especially this is significant, as a smaller range, such as a few seconds difference, could be attributed to mood on the day or other totally unpredictable factors.

### 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)  
  
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/h)", y = "Time (min)",   
 title = "Scatterplot with line of best fit and 95% confidence and prediction intervals")

## `geom\_smooth()` using formula 'y ~ x'



### 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.

predict(Running.model, newdata = data.frame(v4mM = 18.9))

## 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)

These predictions are based off the prediction interval, meaning that they are likely to contain the response variable for a single new observation of a specific value of the exploratory variable. However, the observed running time would still fall within the 95% confidence interval for that value, so it would not imply that the model is incorrect.

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.

predict(Running.model, newdata = data.frame(v4mM = 2.6))

## 1   
## 14.85286

The concern that I would have is that 2.6km/h is not within the range found in the study, and we don’t have any concrete data to base predictions on for that.

### 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.

In conclusion, fitting a linear model to this relationship seems accurate, as there were not any extreme datapoints, there was a very high correlation coefficient, and the majority of datapoints fell within the 95% confidence interval that was predicted.

Therefore it is reasonable to say that a linear regression of this relationship can be used to accurately predict the time taken to run 3km given the velocity attained at a blood lactate concentration of 4mmol/litre.