Simple Linear Regression Assignment

Bryan Mannix

Predicting 3 km Running Times based on laboratory testing

Study Description

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

Aims

To investigate whether there is sufficient evidence of a dependency of 3 km running time on v-4mM in the population of male runners of interest in order to use their blood lactate markers to predict their 3km running time.

```r
# Load the libraries needed.
library(tidyverse)


# read in the data
running.df <- read.csv("3krunning.csv", header = TRUE)
summary(running.df)
```

```
##   Running.Time         v4mM           vTlac          Rel.14.5    
##  Min.   : 8.230   Min.   :14.20   Min.   :13.50   Min.   :46.50  
##  1st Qu.: 9.090   1st Qu.:15.47   1st Qu.:14.55   1st Qu.:49.60  
##  Median : 9.390   Median :17.25   Median :16.00   Median :51.15  
##  Mean   : 9.458   Mean   :17.07   Mean   :15.95   Mean   :51.59  
##  3rd Qu.:10.100   3rd Qu.:18.45   3rd Qu.:17.07   3rd Qu.:53.67  
##  Max.   :10.580   Max.   :20.40   Max.   :19.50   Max.   :57.50  
##     Rel.16.1         VO2Max     
##  Min.   :50.60   Min.   :16.20  
##  1st Qu.:55.75   1st Qu.:19.62  
##  Median :57.45   Median :21.20  
##   Mean   :57.82   Mean   :20.69  
##   3rd Qu.:60.42   3rd Qu.:22.07  
##   Max.   :64.00   Max.   :23.50
```

The hypothesis we are testing is whether blood lactate concentration has an effect on 3 km running time. That is,

$H_0:\beta_{v4mM}=0$H0:βv4mM=0
$H_1:\beta_{v4mM}\neq 0$H1:βv4mM≠0
where β_{v4mM}βv4mM is the slope coefficient of a simple linear regression model of 3 km running time (`Running.Time`) on blood lactate concentration (`v4mM`).

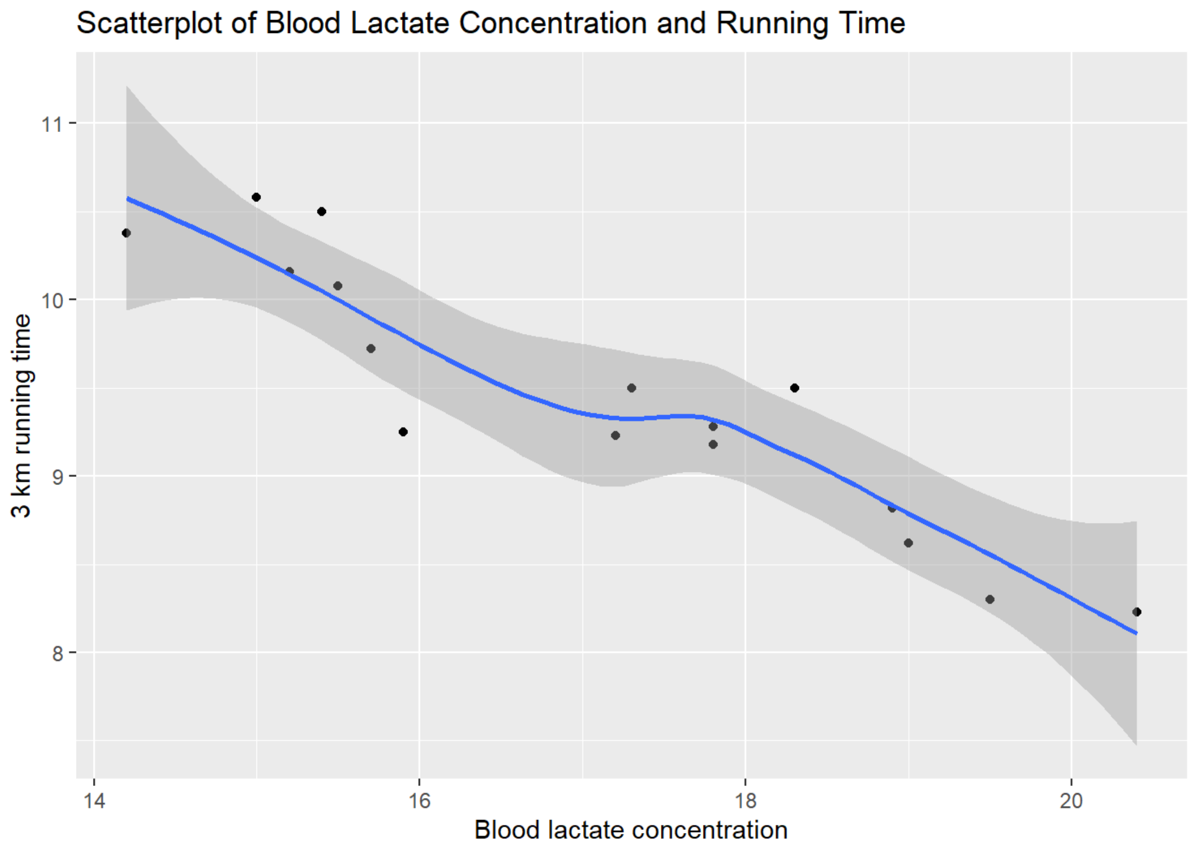
Exploratory Data Analysis

```
running.df %>%
  summarize(Mean.Running.Time = mean(Running.Time),
            SD.Running.Time = sd(Running.Time),
            Mean.v4mM = mean(v4mM),
            S.v4mM = sd(v4mM))
```

```
##   Mean.Running.Time SD.Running.Time Mean.v4mM   S.v4mM
## 1          9.458125        0.744269  17.06875 1.848141
```

```
ggplot(running.df, aes(y = Running.Time, x = v4mM)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Blood lactate concentration", y = "3 km running time",
       title = "Scatterplot of Blood Lactate Concentration and Running Time
")
```

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



The scatterplot suggests that there is a strong negative linear relationship between blood lactate concentration and 3 km running time. That is, higher blood lactate concentrations correspond to faster running times. This is confirmed by the correlation coefficient of -0.926.

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

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

Formal Analysis

```
running.model <- lm(Running.Time ~ v4mM, running.df)
summary(running.model)
```

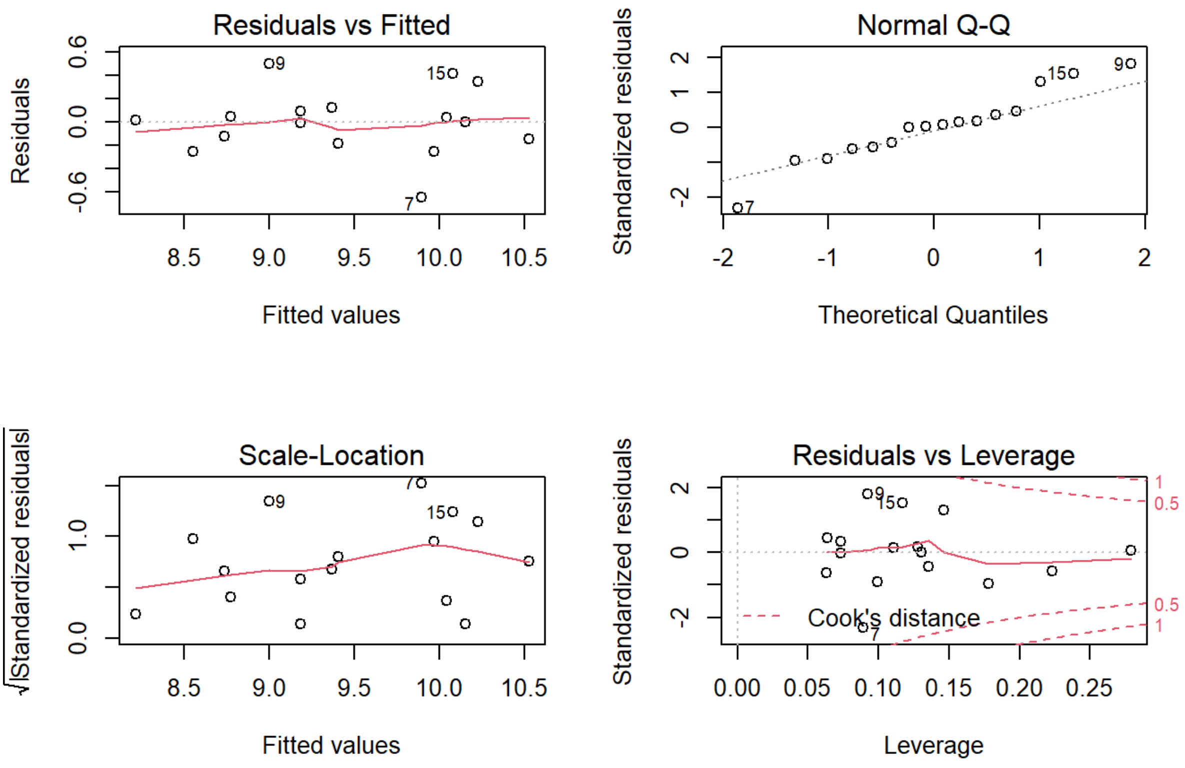
```
##
## Call:
## lm(formula = Running.Time ~ v4mM, data = running.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64390 -0.15561  0.00952  0.10292  0.50095
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 15.82228    0.69800  22.668 1.96e-12 ***
## v4mM        -0.37285    0.04067  -9.168 2.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2911 on 14 degrees of freedom
## Multiple R-squared:  0.8572, Adjusted R-squared:  0.847 
## F-statistic: 84.05 on 1 and 14 DF,  p-value: 2.71e-07
```

The estimated simple linear regression model is

$\hat{Running.Time}=15.82228-0.37285v4mM$Running.Time^=15.82228−0.37285v4mM
The slope coefficient is statistically significant (p-value = 2.71e-07), indicating that blood lactate markers can be used to predict 3km running time in the population of male well-trained middle and long distance runners.

```
par(mfrow = c(2, 2))
plot(running.model)
```



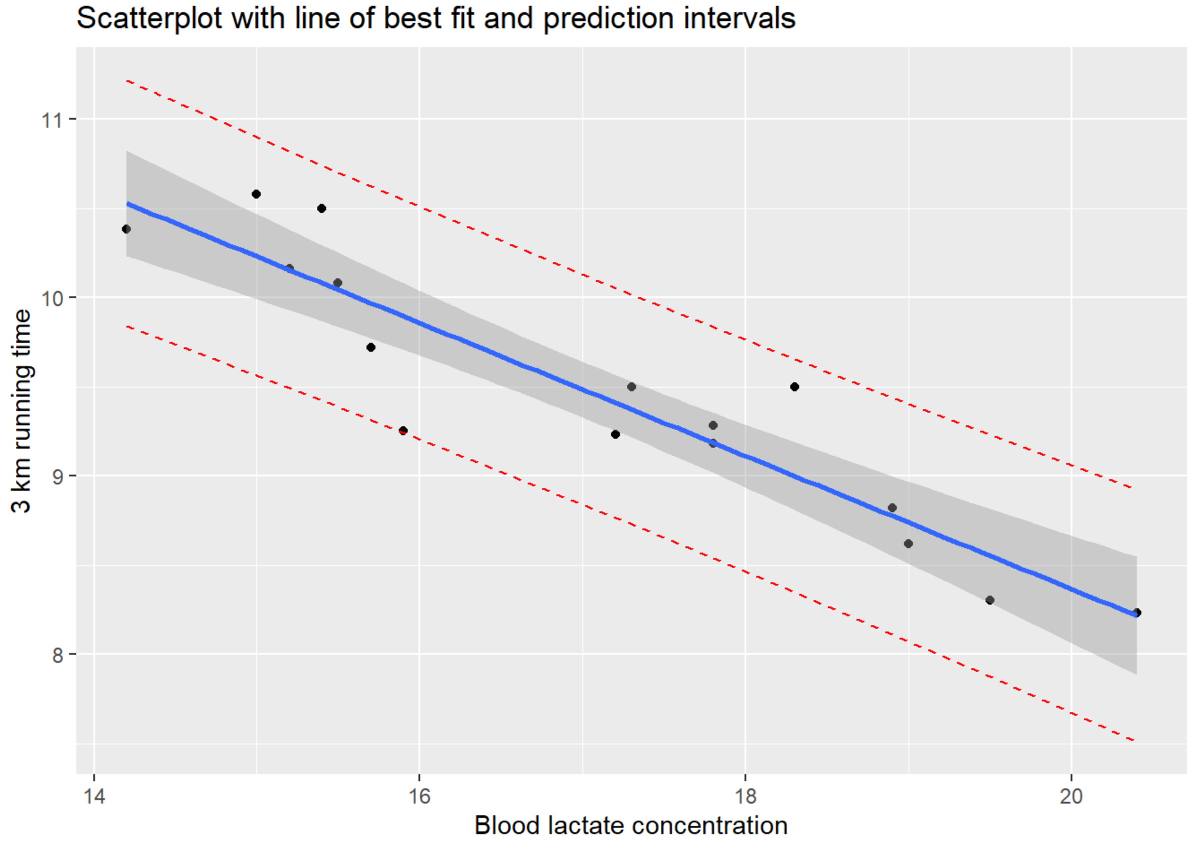
The diagnostic plots show no issues with the model.

```r
pred.int <-  predict(running.model, newdata = running.df, interval = "prediction")


running.df2 <- cbind(running.df, pred.int)


ggplot(running.df2, aes(y = Running.Time, x = v4mM)) +
  geom_point() +
  stat_smooth(method = lm) +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  labs(x = "Blood lactate concentration", y = "3 km running time",
       title = "Scatterplot with line of best fit and prediction intervals")
)
```

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



Conclusion and Translation

There is sufficient evidence of a dependency of 3 km running time on v-4mM in the population of male runners of interest. In particular, each one-unit increase in v-4mM decreases 3 km running time by about 0.37 minutes, on average, and v-4mM explains about 85.72% of variability in 3 km running times. Thus, we can use blood lactate markers to predict the 3km running time of well-trained middle and long distance male runners.