DSCI 561 Lab 2 Solutions

Load all necessary R packages:

```
library("tidyverse", quietly = TRUE)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library("car", quietly = TRUE)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library("knitr", quietly = TRUE)
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

Lab 2 - Exercises on Linear Regression

General instructions

rubric={mechanics:2}

- Ensure your lab meets these basic lab requirements: https://ubc-mds.github.io/resources_pages/general_lab_instructions/
- This assignment is to be completed in R, submitting both a .Rmd markdown file you create in RStudio (you can add your answers directly to this one) along with a rendered `.pdf` **AND** `.md` file (we also want to see a PDF of this lab because of the LaTeX equations).

In Exercises 1 and 2, you will work with the `marathon_full.csv` dataset in our course repository marathon_full.csv Run the following code chunk to extract runners who ran at least a marathon and to create a variable `mf_s`: meters per second.

```
marathon<- read.csv2("./marathonfull.csv", header=TRUE, sep=",")
marathon_ful<- marathon %>%
  filter(cohort1 == 1) %>%
  select(c(age, bmi, female, footwear, group, injury, mf_d, mf_di, mf_ti,  max, sprint, tempo))%>% muta
```

Exercise 1 - Linear regression with 2 discrete explanatory variables

In the dataset, the variable `sprint` indicates whether the runner runs sprints, intervals, or hill repeats most weeks during training (call it sprint trainning); and the variable `tempo` indicates whether the runner does tempo runs most weeks during training (call it tempo trainning). They are both binary variables with 0 = no and 1 = yes.

In general, in terms of their running speed `mf_s`, meters per second, do all runners perform the same, regardless of whether they did sprint trainning? regardless whether they did tempo trainning? Do the effects of sprint trainning depend on tempo trainning? Or vice-versa?

Hint - Start with an interaction term in your model, and ask your hypothesis surrounding this. If the interaction is significant, you can focus on explaining the interaction effect and can disregard interpretating the main effects. If not, then you may have to change your hypothesis, and fit a model without the interaction (or simply disregard the interaction term when interpreting your model).

1A. Understanding the Study Design

rubric={reasoning:3}

- Identify the explanatory and the response variable.
- Write out in words, appropriate null and alternative hypotheses

Solutions

Explanatory variables: whether a runner does sprint trainning (`sprint`) and/or tempo trainning (`tempo`)

Response variable: running speed (`mf_s`)

Hypotheses A

H_0: There is no interaction effect between sprint and tempo trainning on average running speed.

H_1: There is an interaction effect between sprint and tempo trainning on average running speed.

Hypotheses B

H_0: There is no main effect of sprint and tempo trainning on average running speed.

H_1: There are main effects of sprint and tempo trainning on average running speed.

1B. Visualize the data

rubric={reasoning:3}

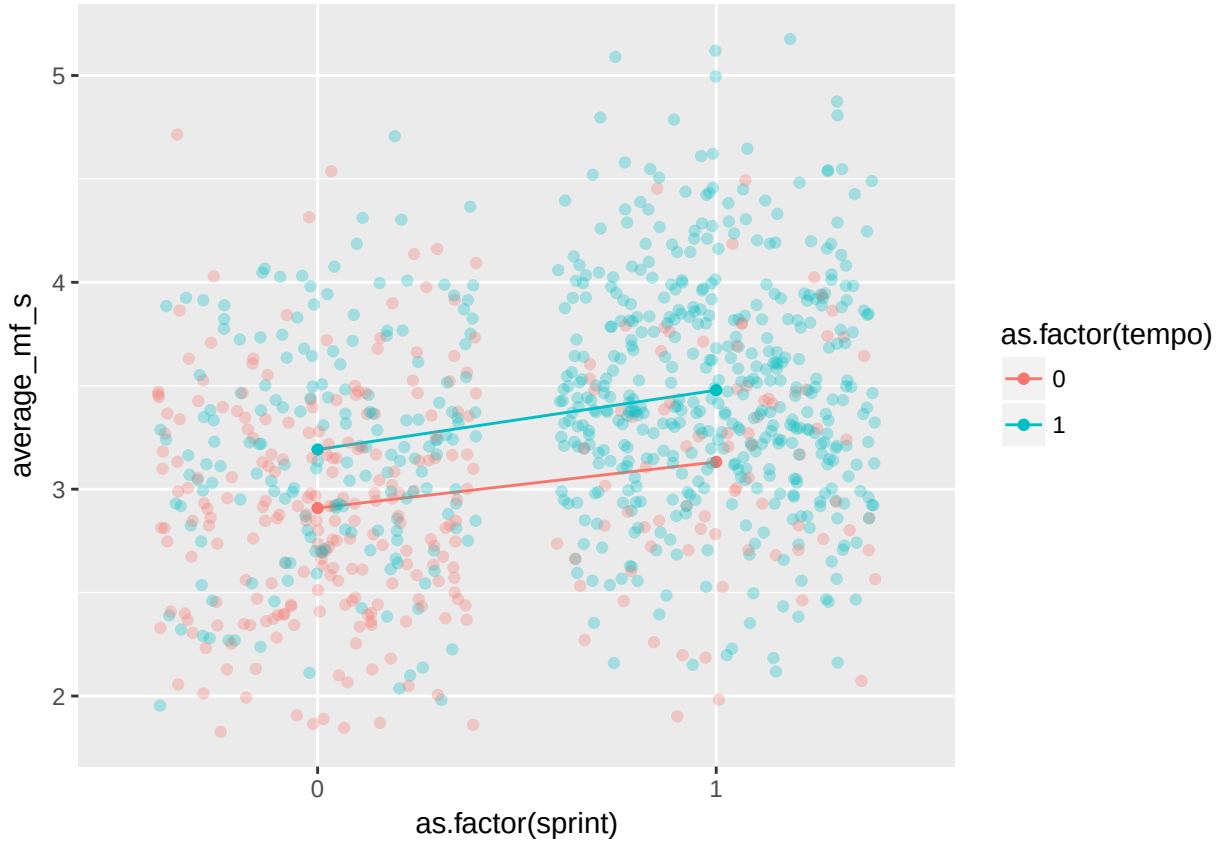
- Use ggplot2's `geom_point`, and `geom_line` to create an interaction plot
- Use ggplot2's `geom_jitter` to visualize the data points on the interaction plot
- Describe what you think the interaction plot is showing you

```
## Extract the explanatory and response variables
Q1_data <- marathon_ful %>% select(sprint,tempo,mf_s)

## compute sample groups mean
```

```
group_sample_mean <-
  Q1_data %>% group_by(sprint,tempo) %>% summarise(average_mf_s=mean(mf_s))

## interaction plot
ggplot(group_sample_mean,aes(x=as.factor(sprint),y=average_mf_s,colour=as.factor(tempo))) +
  geom_point() +
  geom_line(aes(group=as.factor(tempo))) +
  geom_jitter(data = Q1_data,aes(x=as.factor(sprint),y=mf_s),alpha = 0.3)
```



The interaction plot shows two distinct and parallel increasing lines. This implies that sprint and tempo training are likely to have non-zero main effects on running speed but no interaction effect.

1C. Fit a linear model to the data and explain the output of the model in regards to the hypothesis

rubric={reasoning:4}

- Fit a linear model to the data
- Explain the estimated intercept term in the context of the question.
- Explain the model results in the context of the hypothesis.

```
## fit model with interaction effect
lm_with_interaction <- lm(data = Q1_data,formula = mf_s~sprint*tempo)
summary(lm_with_interaction)
```

```
##
## Call:
## lm(formula = mf_s ~ sprint * tempo, data = Q1_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3602 -0.3679 -0.0440  0.3644  1.8057
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.90883    0.03659  79.488  < 2e-16 ***
## sprint        0.22290    0.07081   3.148   0.0017 **
## tempo         0.28287    0.05506   5.137  3.4e-07 ***
## sprint:tempo  0.06386    0.08594   0.743   0.4576
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5489 on 925 degrees of freedom
## Multiple R-squared:  0.1549, Adjusted R-squared:  0.1522
## F-statistic: 56.52 on 3 and 925 DF,  p-value: < 2.2e-16
```

We begin with examining the interaction effect between sprint and tempo training on running speed using a linear model with interaction term, i.e., examining the hypothese A. The associated p-value for the interaction effect is 0.4576 so there is no significant evidence that thereis a non-zero interaction effect. We then focus on interpreting the main effects of the two training in a second linear model without the interaction term.

```
## fit model without interaction effect
lm_without_interaction <- lm(data = Q1_data,formula = mf_s~sprint+tempo)
summary(lm_without_interaction)
```

```
##
## Call:
## lm(formula = mf_s ~ sprint + tempo, data = Q1_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35438 -0.37458 -0.03904  0.37041  1.81727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.89725    0.03310  87.523  < 2e-16 ***
## sprint       0.26626    0.04011   6.637 5.43e-11 ***
## tempo        0.30909    0.04226   7.313 5.64e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5488 on 926 degrees of freedom
## Multiple R-squared:  0.1544, Adjusted R-squared:  0.1526
## F-statistic: 84.55 on 2 and 926 DF,  p-value: < 2.2e-16
```

In this linear model, the estimated intercept term represents the estimated mean running speed of runners who do not do any of the two trainings.

The result indicates that both the sprint and tempo training have significant positive main effects on running speed. This implies that runners who do sprint and/or tempo training are likely to have a higher mean running speed than runners who do not, and runners who do both of the trainings are likely to have a higher mean running speed than runners who do only one of them.

1D. Comparing ANOVA and linear regression

rubric={reasoning:3}

- Use `aov()` to perform an two-way analysis of variance (ANOVA) in R to compare the running speed `mf_s` among the four groups with different sprint training and tempo training status. Use `summary()` to produce the results from your model object.
- Discuss the results you obtained from applying the different methods (lm in Part 1C and anova in Part 1D) to the same question. Do you get the same results or different? Is this what you expected? Why or why not?

```
aov_with_interaction <- aov(data = Q1_data,formula = mf_s~sprint*tempo)
summary(aov_with_interaction)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## sprint         1  34.82   34.82 115.561  < 2e-16 ***
## tempo          1  16.11   16.11  53.457 5.72e-13 ***
## sprint:tempo   1   0.17    0.17   0.552    0.458
## Residuals    925 278.72    0.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> The result is exactly the same as the result of the linear model with interaction term in Part 1C. This is what we expected since both the linear model with factor explanatory variable and ANOVA are assessing the existence of difference in mean running speed between the four groups of runners receiving different combinations of trainings.

Exercise 2 - linear regression with 1 continuous explanatory variable

The variable `max` indicates the maximum number of miles the runner ran in a single week during training. Is there a relationship between `max` and their running speed `mf_s` in the competitions?

2A. Understanding the Study Design

rubric={reasoning:3}

```
- Identify the explanatory and the response variables.
- Write out in words, appropriate null and alternative hypothesess
```

Solutions

```
#looking at some data values
marathon_ful %>% select(max, mf_s) %>% some()
```

```
##     max     mf_s
## 55   60 2.849666
## 84   55 3.396523
## 222  55 3.001067
## 377  45 2.990644
## 469  62 3.497306
## 480  84 3.286983
## 497  40 3.471982
## 519  50 2.813000
## 780  48 2.756042
## 884  50 2.706369
```

Explanatory variable: maximum number of miles the runner ran (`max`)

Response variable: running speed (`mf_s`)

H_0: There is no linear relationship between `max` and `mf_s`.

H_A: There is a linear relationship between `max` and `mf_s`.

2B. Visualize the data

rubric={reasoning:2}

- Use ggplot2's `geom_point`, and to create an scatter plot of the data
- Describe what you think the scatter plot is showing you

Solutions

```
dat <- marathon_ful %>% select(max, mf_s) %>% mutate(maxn=as.numeric(max))
ggplot(dat, aes(x=maxn, y=mf_s))+geom_point() + xlab("Single week training (maximum miles ran)") + ylab
```

The outliers with fast runners that didn't train as much will skew the regression line. However, there appears to be a positive linear relationship between `max` and `mf_s`.

2C. Fit a linear model to the data

rubric={reasoning:2}

- Fit a linear model to the data

Solutions

```
fit <- lm(mf_s~maxn, data=dat)
summary(fit)
```

```
##
## Call:
## lm(formula = mf_s ~ maxn, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45281 -0.36440 -0.03261  0.33530  2.38849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.705198   0.067686  39.967   <2e-16 ***
## maxn        0.006334   0.000749   8.457   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5747 on 927 degrees of freedom
## Multiple R-squared:  0.07163,    Adjusted R-squared:  0.07062
## F-statistic: 71.52 on 1 and 927 DF,  p-value: < 2.2e-16
```

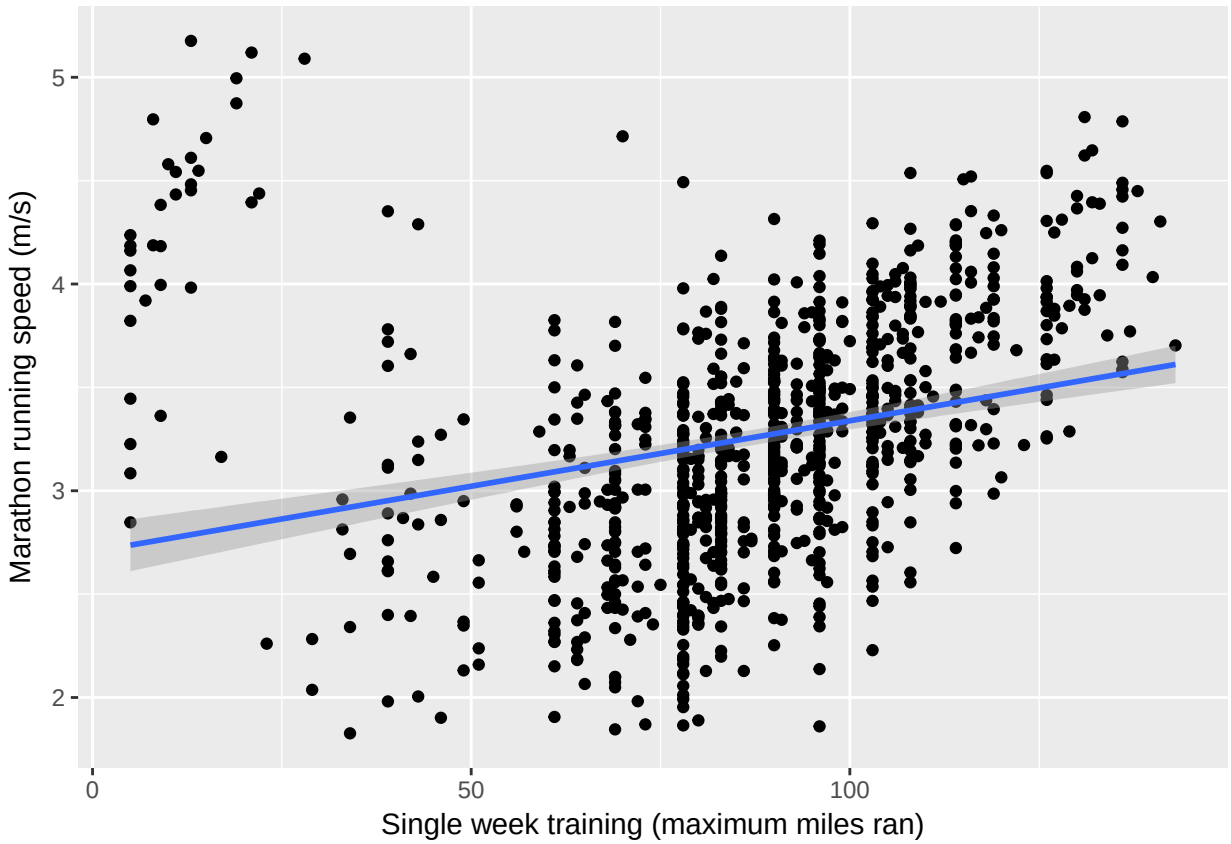
2D. Visualize the model

rubric={reasoning:3}

- Use ggplot2's geom_point() + geom_smooth(method = "lm") to overlay the regression line over the scatter plot of the data
- Based on this visualization, discuss how well do you think the model fits the data

Solutions

```
ggplot(dat, aes(x=maxn, y=mf_s))+geom_point()+geom_smooth(method='lm') + xlab("Single week training (ma
```



As mentioned, the outliers increased β_0 and decreased β_1. The R^2 is the square of the correlation between response and the fitted values. It suggests the data is not well fitted by the regression model.

Remving the outliers we get a better model. R^2 is now 0.34 and β_0 and β_1 aren't skewed anymore.

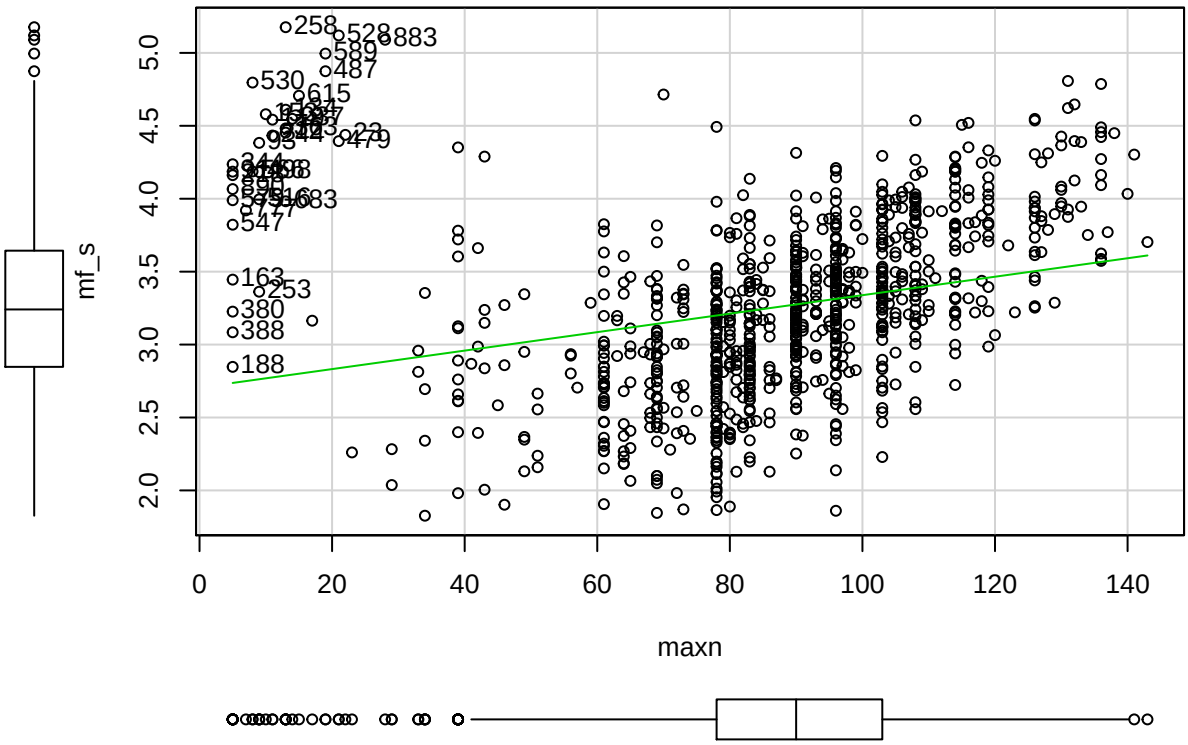
```
nrow(dat)
```

```
## [1] 929
```

```
scatterplot(mf_s ~ maxn, data=dat, smooth=FALSE, id.n=33)
```

```
##  12  23  91  93 119 134 153 163 188 218 237 244 253 258 344 363 380 388
##  12  23  91  93 119 134 153 163 188 218 237 244 253 258 344 363 380 388
## 479 487 498 516 528 530 547 566 579 589 615 683 777 883 890
## 479 487 498 516 528 530 547 566 579 589 615 683 777 883 890
```

```
out <- scatterplot(mf_s ~ maxn, data=dat, smooth=FALSE, id.n=33)
```




```
updated.fit <- update(fit, subset=-out)
summary(updated.fit)
```

```
##
## Call:
## lm(formula = mf_s ~ maxn, data = dat, subset = -out)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46000 -0.30582  0.01305  0.30811  1.93954
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.7908597  0.0670144   26.72   <2e-16 ***
## maxn        0.0159332  0.0007285   21.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4557 on 894 degrees of freedom
## Multiple R-squared:  0.3486, Adjusted R-squared:  0.3478
## F-statistic: 478.4 on 1 and 894 DF,  p-value: < 2.2e-16
```

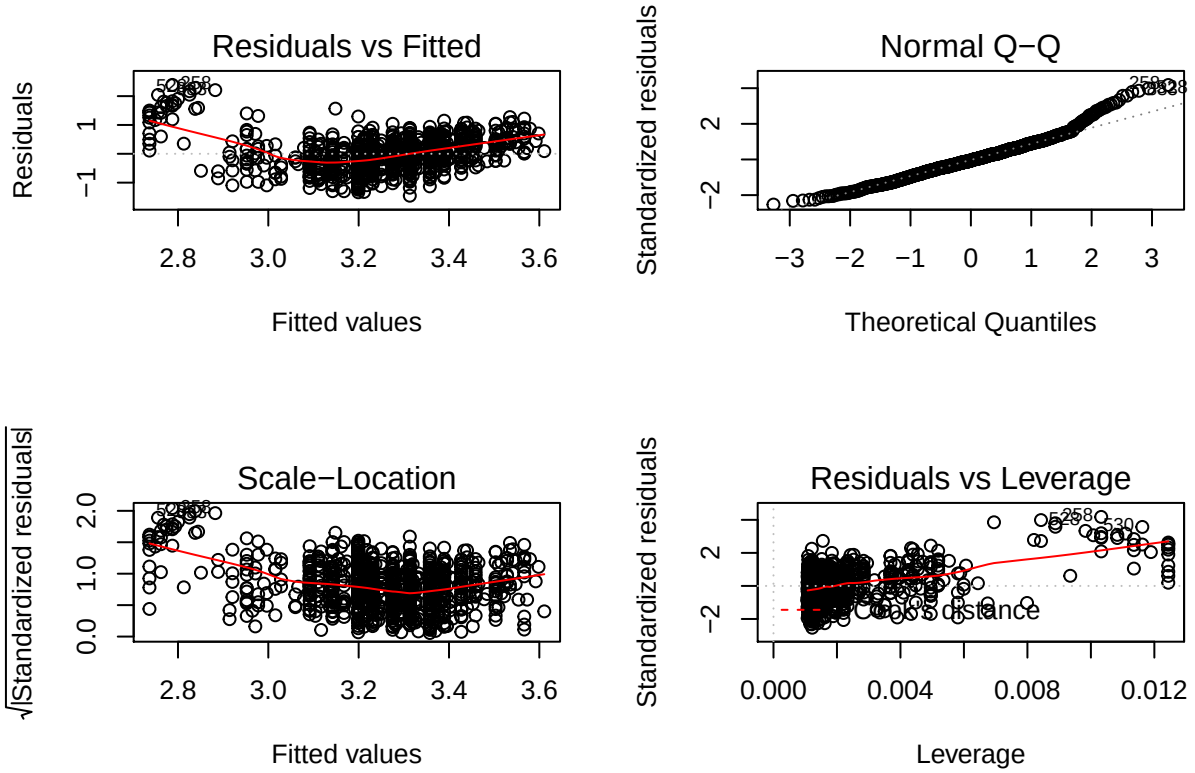
2E. Explain the output of the model in regards to the hypothesis

rubric={reasoning:4}

- Explain the model results in the context of the hypothesis
- What assumptions are you making for the model?

Solutions

```
par(mfrow=c(2,2))
plot(fit)
```



Given the small p-value of the slope, we have enough evidence to reject the null hypothesis and conclude that there is a relationship between maximum number of miles the runner ran in a single week during training, and their running speed in the marathon.

The QQ plot shows the errors are Normally distributed. However, there is a pattern in the residual plot. This plot allows us to evaluate the assumption of constant variance in the residuals and check if the plot has some pattern. We observe *heteroscedicity* (non constant variance). Our constant variance assumption has been violated. This means we should consider a different model.

Exercise 3 - Linear regression with many discrete and continuous explanatory variables, as well as an interaction term

From `marathon_full.csv` data set, come up with a question you are interested in asking, such that you need to perform a linear regression with >= 2 discrete and >= 2 continuous explanatory variables.

3A. Understanding the Study Design

rubric={reasoning:2}

```
- Identify the explanatory and the response variable.
- Write out in words, appropriate null and alternative hypotheses
```

Discrete explanatory variables: 'sprint', 'tempo'

Continuous explanatory variables: 'bmi', 'max'

Response variable: 'mf_s'

H_0: The linear relationship between `max` and `mf_s` is not affected by whether the runner does sprint and/or tempo traning ('sprint', 'tempo') controlling for the 'bmi'

H_1: The linear relationship between `max` and `mf_s` is affected by whether the runner does sprint and/or tempo traning ('sprint', 'tempo') controlling for the 'bmi'

3B. Visualize the data

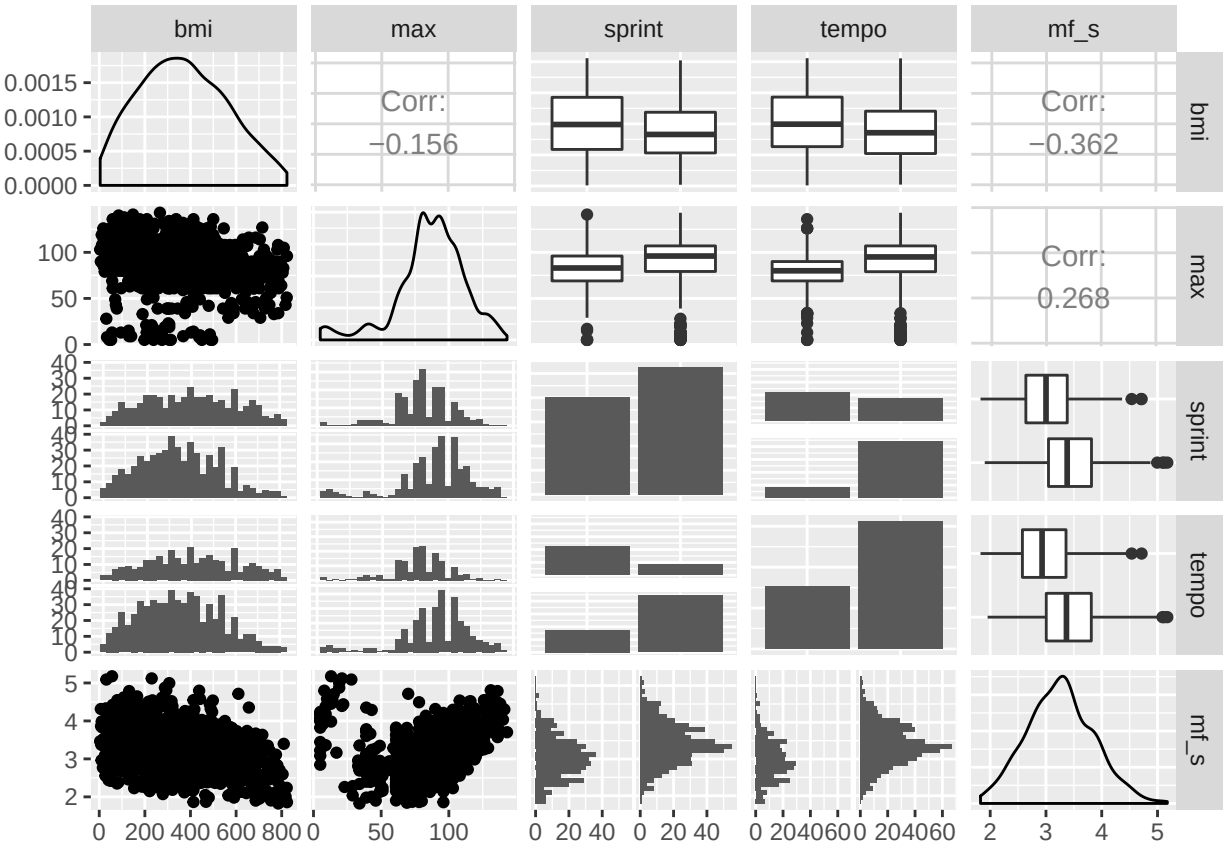
rubric={reasoning:2}

```
- Use `GGally::ggpairs()` to make a pair plot to visualize the data
- Describe what you think the plot is showing you
```

```
## Extract data
Q3_data <- marathon_ful %>% select(bmi,max,sprint,tempo,mf_s)
Q3_data$bmi <- as.numeric(Q3_data$bmi)
Q3_data$max <- as.numeric(Q3_data$max)
Q3_data$sprint <- as.factor(Q3_data$sprint)
Q3_data$tempo <- as.factor(Q3_data$tempo)

ggpairs(data = Q3_data)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot tells us that

a. There is a positive linear relationship between 'max' and 'mf_s'
b. There is a negative linear relationship between 'bmi' and 'mf_s'
c. Both 'sprint' and 'tempo' have positive effect on 'mf_s'
d. There is a negative linear relationship between 'bmi' and 'max'

3C. Fit a linear model to the data

rubric={reasoning:2}

```
- Fit a linear model to the data
```

```
lm_Q3 <- lm(data = Q3_data, formula = mf_s ~ sprint*max + tempo*max + bmi)
summary(lm_Q3)
```

```
##
## Call:
## lm(formula = mf_s ~ sprint * max + tempo * max + bmi, data = Q3_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41285 -0.35409 -0.03229  0.32307  2.06148
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.573e+00  1.314e-01  19.580  < 2e-16 ***
## sprint1      7.190e-01  1.403e-01   5.125 3.63e-07 ***
## max          8.660e-03  1.496e-03   5.789 9.71e-09 ***
## tempo1       3.413e-01  1.503e-01   2.271 0.023367 *
## bmi         -8.632e-04  8.928e-05  -9.669  < 2e-16 ***
## sprint1:max -5.903e-03  1.584e-03  -3.726 0.000206 ***
## max:tempo1  -1.396e-03  1.762e-03  -0.792 0.428391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5072 on 922 degrees of freedom
## Multiple R-squared:  0.281,  Adjusted R-squared:  0.2763
## F-statistic: 60.04 on 6 and 922 DF,  p-value: < 2.2e-16
```

3D. Explain the output of the model in regards to the hypothesis

rubric={reasoning:2}

```
- Explain the model results in the context of the hypothesis
```

The fitted model shows a significant weaker linear relationship between 'max' and 'mf_s' if the runner does sprint training compare to runner does not do any of the two trainings (sprint1:max, estimate: -5.903e-03, p-value: 0.000206); though there is no significant different in the relationship between 'max' and 'mf_s' between runner does tempo training and runner does not do any of the two trainings (max:tempo1, estimate: -1.396e-03, p-value: 0.428391).