DSCI 561 Lab 1 Solutions

Lab 1 - Intro to Linear Regression

Load all necessary R packages:

```r
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
```

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

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

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

```r
library("broom", quietly = TRUE)
library("Lahman", quietly = TRUE)
library("reshape", quietly = TRUE)
```

```
##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
##     rename

## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

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

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

## The following object is masked from 'package:Lahman':
##
##     Salaries

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

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

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

Lab 1 - Intro to Linear Regression

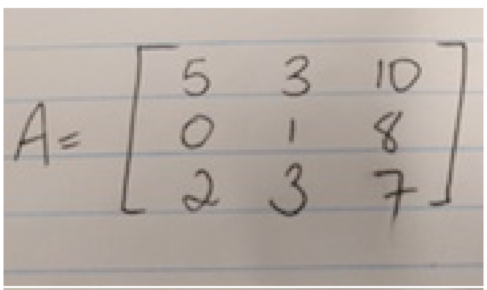
General instructions

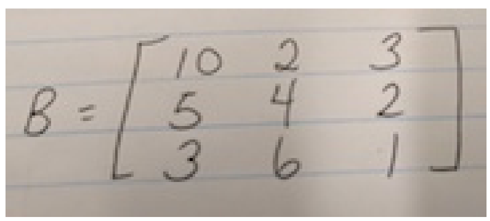
rubric={mechanics:2}

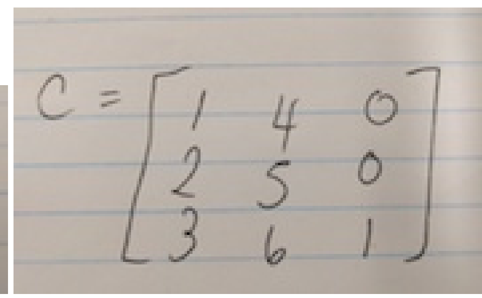
- This assignment is to be completed in R, submitting both a .Rmd markdown file you create in RStudio (you can add your answers directly to this one) along with a rendered `.pdf` **AND** `.md` file (we also want to see a PDF of this lab because of the LaTeX equations).

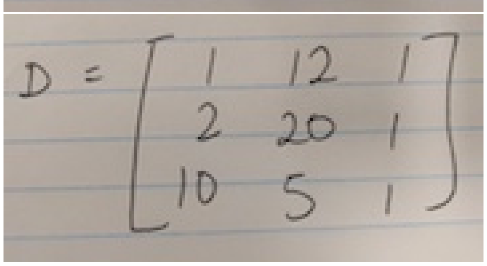
Exercise 0 - Reminder of matrix notation and algebra

Use the four matrices below to answer the questions in this exercise:









hint - this link provides basic examples of Matrix Algebra: http://stattrek.com/matrix-algebra/matrix-addition.aspx

0A.

rubric={reasoning:1}

If $E = A + B$, what is e_{23} ?

0B.

rubric={reasoning:1}

Find the matrix BC

0C.

rubric={reasoning:1}

- Find the inverse of D
- You can use the `solve()` function to do this.

0D.

rubric={reasoning:1}

Provide an example of a where $XY = YX$ and an example where $XY \neq YX$

Solutions

```
##0A

#Define matrices:
A <- matrix(c(5,3,10,0,1,8,2,3,7),nrow = 3,ncol = 3, byrow = T)
B <- matrix(c(10,2,3,5,4,2,3,6,1),nrow = 3,ncol = 3, byrow = T)
C <- matrix(c(1,4,0,2,5,0,3,6,1),nrow = 3,ncol = 3, byrow = T)
D <- matrix(c(1,12,1,2,20,1,10,5,1),nrow = 3,ncol = 3, byrow = T)

A
```

```
##      [,1] [,2] [,3]
## [1,]    5    3   10
## [2,]    0    1    8
## [3,]    2    3    7
```

```
B
```

```
##      [,1] [,2] [,3]
## [1,]   10    2    3
## [2,]    5    4    2
## [3,]    3    6    1
```

```
E <- A + B

E
```

```
##      [,1] [,2] [,3]
## [1,]   15    5   13
## [2,]    5    5   10
## [3,]    5    9    8
```

```
## 0B.

#Multiply matrices B and C

B
```

```
##      [,1] [,2] [,3]
## [1,]   10    2    3
## [2,]    5    4    2
## [3,]    3    6    1
```

```
C
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    0
```

```
## [2,]    2    5    0
## [3,]    3    6    1
```

```
BC = B %*% C
BC
```

```
##      [,1] [,2] [,3]
## [1,]   23   68    3
## [2,]   19   52    2
## [3,]   18   48    1
```

```
##0C.

#inverse of D
invD = solve(D)
invD
```

```
##            [,1]        [,2]        [,3]
## [1,] -0.1898734  0.08860759  0.10126582
## [2,] -0.1012658  0.11392405 -0.01265823
## [3,]  2.4050633 -1.45569620  0.05063291
```

```
##0D.

# matrices B and C are not commutative
BC
```

```
##      [,1] [,2] [,3]
## [1,]   23   68    3
## [2,]   19   52    2
## [3,]   18   48    1
```

```
CB = C %*% B
CB
```

```
##      [,1] [,2] [,3]
## [1,]   30   18   11
## [2,]   45   24   16
## [3,]   63   36   22
```

```
# The identity matrix is commutative:
I <- matrix(c(1,0,0,0,1,0,0,0,1),nrow = 3,ncol = 3, byrow = T)
I
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```
IB = I %*% B
IB
```

```
##      [,1] [,2] [,3]
## [1,]   10    2    3
## [2,]    5    4    2
## [3,]    3    6    1
```

```
BI = B %*% I
BI
```

```
##      [,1] [,2] [,3]
```

```
## [1,]   10    2    3
## [2,]    5    4    2
## [3,]    3    6    1
```

Exercise 1 - Exploring data with a single discrete explanatory variable that is composed of two groups

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))%>% mutate(mf_s
```

In the dataset, the variable `footwear` indicates what type of footwear the runners wear:

- 1 = Minimalist
- 2 = Normal running shoe
- 3 = Vibrams, sandals, or barefoot

In general, do runners wearing `footwear == Minimalist` perform differently than runners wearing `footwear == Normal running shoe` in terms of their running speed `mf_s`, meters per second.

We can extract samples from the dataset to answer this question.

```
marathon_ful %>% filter(footwear==1) %>% some()
```

```
##     age         bmi female footwear group injury  mf_d mf_di mf_ti max
## 41   30 23.74768066      0        1     3      1 42195     4  9278 112
## 45   28 22.68170738      0        1     2      1 42195     4  9615  88
## 46   32 19.34570503      0        1     1      1 42195     4  8152 110
## 54   47 21.96660423      0        1     1      1 42195     3 13162  55
## 67   41 24.54295158      0        1     1      1 42195     3  9960 100
## 75   37 23.88304138      0        1     1      1 42195     3 11987  40
## 120  54 22.33406639      1        1     3      1 42195     3 12649  65
## 121  32 26.16460609      1        1     3      1 42195     3 19200  45
## 128  36 21.38588142      1        1     1      1 42195     2 17310  40
## 170  28  20.9185257      0        1     1      2 42195     2 10717  63
##     sprint      mf_s
## 41       1 4.547855
## 45       1 4.388456
## 46       1 5.176030
## 54       0 3.205820
## 67       1 4.236446
## 75       1 3.520063
## 120      0 3.335837
## 121      1 2.197656
## 128      0 2.437608
## 170      1 3.937203
```

```
marathon_ful %>% filter(footwear==2) %>% some()
```

```
##     age         bmi female footwear group injury  mf_d mf_di mf_ti max
## 25   27 20.77414703      1        2     2      1 42195     2 12095  70
## 64   31 23.30241394      1        2     3      3 42195     3 12123  50
```

```
## 188  42 23.00573349      1         2     3       1 42195      3 14820  48
## 304  48 25.81369209      0         2     2       2 42195      3 13835  56
## 341  38  23.5923233      0         2     3       2 42195      3 12120  55
## 366  36 21.51694489      1         2     2       1 42195      3 11400  35
## 397  49   20.522686      1         2     2       1 42195      4 13020  65
## 528  34 23.00556564      0         2     3       1 42195      4 10750  75
## 565  30 22.44668961      0         2     1       2 42195      3 13200  30
## 600  36 23.16774559      1         2     2       1 42195      2 17101  60
##     sprint     mf_s
## 25       1 3.488632
## 64       1 3.480574
## 188      0 2.847166
## 304      1 3.049874
## 341      0 3.481436
## 366      0 3.701316
## 397      0 3.240783
## 528      0 3.925116
## 565      1 3.196591
## 600      1 2.467400
```

1A. Understanding the Study Design

rubric={reasoning:3}

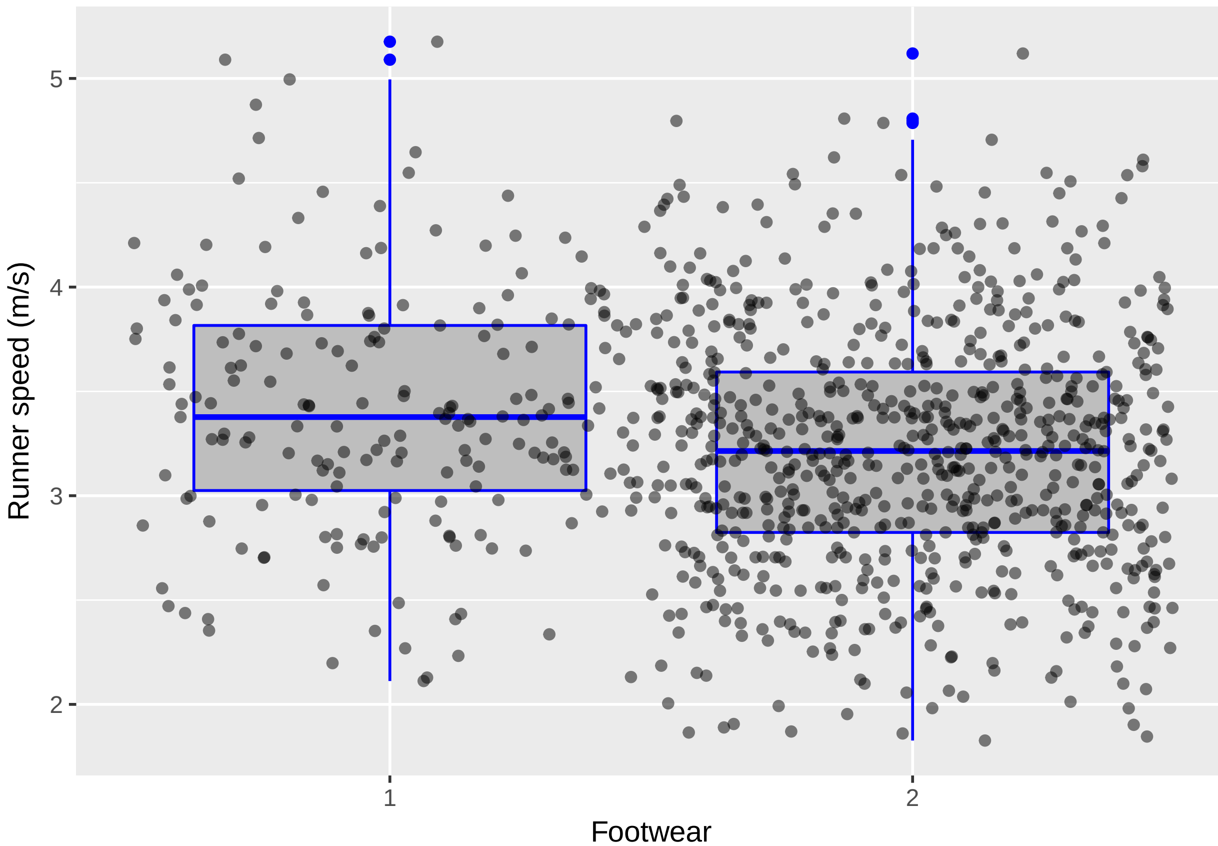
- Identify the explanatory and the response variable.
- Write out in words, appropriate null and alternative hypotheses
- Create a boxplot of the data to compare the two groups
- Describe the graph, for example:
  - does it look as if the groups have equal means or equal variance?
  - are there any unusual observations in the data set?
- Calculate the mean, number of observations and standard deviation for each of the groups

Solutions

> Explanatory variable: footwear
>
> Response variable: running speed
>
> H_0: runners wearing `footwear == Minimalist` and runners wearing `footwear == Normal running shoe`have the same population mean running speed
>
> H_a: runners wearing `footwear == Minimalist` and runners wearing `footwear == Normal running shoe`have different population mean running speeds

```
# get data for marathon_ful
marathon_reduced <- marathon_ful %>% filter(footwear!=3)

# visualize data as a boxplot
ggplot(marathon_reduced, aes(as.factor(footwear), mf_s)) + geom_boxplot(fill="grey", colour="blue") + ge
```



It seems that the groups have different means but similar variance. The are several outliers with high running speed in both groups.

```
marathon_reduced_summary <- marathon_reduced %>%
  group_by(footwear) %>%
  summarise(mean_speed = mean(mf_s), n = n(), sd_spped = sd(mf_s))
kable(marathon_reduced_summary)
```

footwear	mean_speed	n	sd_spped
1	3.411465	199	0.590534
2	3.216323	717	0.589657

1B. Comparing the groups means via a t-test, ANOVA and linear regression

rubric={reasoning:2}

- Use `t.test()` with to perform a two-sample t-test in R to compare the groups.
- Use `lm()` to create a model object for your comparison (e.g., `lm(y ~ x, data = df)`)
- Use `anova()` to perform an analysis of variance (ANOVA) in R on your model object to compare the groups.
- Use `summary()` to produce the results from your model object to perform the linear regression in R to compare the groups.
- Explore `broom::tidy()` to get the results from all the tests above.
- Report the results obtained in a nice data frame or table

Solutions

```r
# t-test
ttest_result <- tidy(t.test(marathon_reduced$mf_s~marathon_reduced$footwear,var.equal=TRUE))
kable(ttest_result)
```

estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
3.411465	3.216323	4.129036	3.98e-05	914	0.1023893	0.2878938	Two Sample t-test	two.sided

```r
# create lm model object
lm1 <- lm(data = marathon_reduced,formula = mf_s~footwear)
```

```r
## ANOVA on lm object created
anova_result <- tidy(anova(lm1))
kable(anova_result)
```

term	df	sumsq	meansq	statistic	p.value
footwear	1	5.93166	5.9316603	17.04894	3.98e-05
Residuals	914	317.99855	0.3479196	NA	NA

```r
## summary on lm object created
lm_result <- tidy(summary(lm1))
kable(lm_result)
```

term	estimate	std.error	statistic	p.value
(Intercept)	3.6066062	0.0864789	41.705035	0.00e+00
footwear	-0.1951416	0.0472608	-4.129036	3.98e-05

```r
# data frame with results
kable(data.frame(method = c("t-test", "ANOVA", "Linear regression"),
          p_value = c(ttest_result$p.value, anova_result$p.value[1], lm_result$p.value[2])))
```

method	p_value
t-test	3.98e-05
ANOVA	3.98e-05
Linear regression	3.98e-05

1C. What can we conclude from applying these different methods to this case?

rubric={reasoning:2}

Discuss the results you obtained from applying the different methods in Part B to the same question. Do you get the same results or different? Is this what you expected? Why or why not?

Solutions

> We get the same results from the three different methods (t-test, ANOVA and linear regression). The two sample t-test compares the means of two populations. One-way ANOVA, it compares the means of K groups, which is one factor with K levels. It is asking if the population mean of any of the groups (represented by the one factor of K levels) are different. If $K = 2$ (as in our case), it is equivalent to a two-sample t-test.

ANOVA analysis and linear regression in this context are the same model but presented in different ways. An ANOVA reports a single p-value for the test of the null hypothesis that population mean of the groups (defined by the factor team) are the same. In contrast, linear regression models report the mean for the reference group (specified as an intercept in the model output), and the coefficient represents the estimated difference between this group and the reference group. In the case of the linear model, the p-value represents the probability that we observe the current value of the test statistic for the coefficient estimate or more extreme values under the assumption of the null hypothesis that the difference in population means between groups is zero .

Exercise 2 - Exploring data with a single discrete explanatory variable that is composed of more than two groups

In general, do footwear affect the performance? In other words, do runners wearing `footwear == Minimalist`, `footwear == Normal running shoe` and `footwear = Vibrams, sandals, or barefoot` perform differently in terms of running speed `mf_s`, meters per second.

We can extract samples from the `marathon_ful` dataset to answer this question.

```
marathon_ful %>% filter(footwear==1) %>% some()
```

```
##     age         bmi female footwear group injury  mf_d mf_di mf_ti
## 7    31 22.53840446      0        1     2      2 42195     3 12323
## 71   36 23.67722511      0        1     3      1 42195     3 14712
## 77   39 22.36208534      0        1     1      1 42195     3 10922
## 109  36  23.5923233      0        1     1      2 42195     4 10050
## 116  28 23.08344269      0        1     3      1 42195     3 13320
## 147  30 25.11189079      0        1     2      1 42195     2 11280
## 154  33 22.08595085      1        1     3      3 42195     3 12867
## 159  40 23.84960938      0        1     2      1 42195     3 11100
## 175  29 23.59925842      0        1     2      1 42195     2 10782
## 199  25 19.70114517      1        1     3      1 42195     3 17940
##             max sprint     mf_s
## 7            45      1 3.424085
## 71  20.60000038      1 2.868067
## 77           54      0 3.863303
## 109          70      1 4.198507
## 116          46      0 3.167793
## 147          50      1 3.740691
## 154          40      0 3.279319
## 159          55      1 3.801351
## 175          68      1 3.913467
## 199          42      1 2.352007
```

```
marathon_ful %>% filter(footwear==2) %>% some()
```

```
##     age         bmi female footwear group injury  mf_d mf_di mf_ti max
## 11   39 29.73737335      0        2     1      2 42195     4 13113  52
## 22   26 25.25252533      1        2     1      2 42195     2 16899  34
## 366  36 21.51694489      1        2     2      1 42195     3 11400  35
## 379  32  25.8121376      0        2     2      2 42195     2 14459  43
## 396  35 24.04452515      0        2     3      1 42195     3 10861  74
## 429  48 20.76124382      1        2     1      1 42195     2 12072  55
## 470  26 22.19460106      1        2     2      1 42195     3 15600  37
## 487  33 25.96887398      0        2     2      1 42195     3 12510  47
## 586  35 21.70465088      1        2     1      1 42195     3 13484  50
```

```
## 683  40  23.5923233      0        2     3      3 42195     3 11970  50
##     sprint     mf_s
## 11       0 3.217799
## 22       1 2.496893
## 366      0 3.701316
## 379      0 2.918252
## 396      0 3.885001
## 429      1 3.495278
## 470      1 2.704808
## 487      1 3.372902
## 586      0 3.129264
## 683      1 3.525063
```

```
marathon_ful %>% filter(footwear==3) %>% some()
```

```
##    age         bmi female footwear group injury  mf_d mf_di mf_ti max
## 1   47 25.24751663      0        3     1      1 42195     2 10757  60
## 2   35 25.16514397      1        3     2      1 42195     4 14453  35
## 3   33 21.23309135      0        3     2      1 42195     2 11440  65
## 4   30 21.91380501      0        3     3      3 42195     2 18720  40
## 5   33 25.11223602      0        3     1      1 42195     2 12555  55
## 7   49 23.27272797      0        3     1      1 42195     3 15873  20
## 8   32 19.37938118      0        3     2      1 42195     3 12158  35
## 9   30 24.20903206      0        3     1      1 42195     3 20591  35
## 10  37 28.05836296      1        3     1      3 42195     3 19326  40
## 12  46 22.17678452      0        3     1      1 42195     2 12702  37
##    sprint     mf_s
## 1       1 3.922562
## 2       1 2.919463
## 3       1 3.688374
## 4       0 2.254006
## 5       1 3.360812
## 7       1 2.658288
## 8       0 3.470554
## 9       0 2.049196
## 10      1 2.183328
## 12      0 3.321918
```

2A. Understanding the Study Design

rubric={reasoning:3}

- Identify the explanatory and the response variable.
- Write out in words, appropriate null and alternative hypotheses
- Create a boxplot of the data to compare the three groups
- Describe the graph, for example:
  - does it look as if the groups have equal means or equal variance?
  - are there any unusual observations in the data set?
- Calculate the mean, number of observations and standard deviation for each of the groups

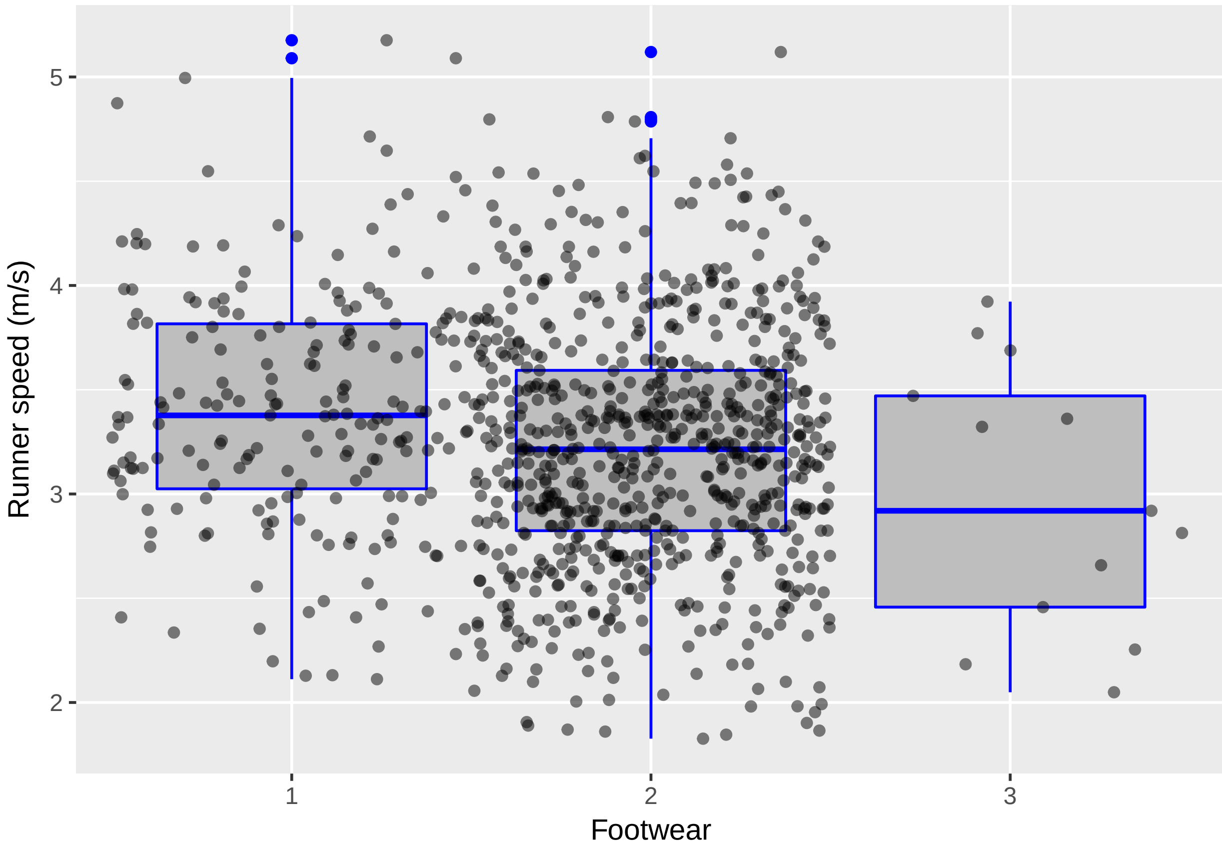
Solutions

explanatory variable: footwear response variable: running speed

H0: There is no difference in runner's mean speed due to differences in footwear HA: There is a difference in runner's mean speed due to differences in at least one type of footwear

```
marathon_ful <- marathon %>%
  filter(cohort1 == 1) %>%
  select(c(age, bmi, female, footwear, group, injury, mf_d, mf_di, mf_ti,  max, sprint)) %>%
  mutate(mf_s = mf_d/mf_ti)


# visualize data as a boxplot
ggplot(marathon_ful, aes(x=as.factor(footwear), y=mf_s))  +
geom_boxplot(fill="grey", colour="blue") + geom_jitter(width = 0.5, alpha = 0.5) + xlab("Footwear") + yl
```



Most runners ran with normal runner shoes. The variance for footwear 3 is slightly higher than for footwear 1 and 2, but the difference is not so large that we can't assume equal variances.

```
marathon_ful_summary <- marathon_ful %>%
  group_by(footwear) %>%
  summarise(mean_speed = mean(mf_s), n = n(), sd_spped = sd(mf_s))

kable(marathon_ful_summary)
```

footwear	mean_speed	n	sd_spped
1	3.411465	199	0.590534
2	3.216323	717	0.589657
3	2.989971	13	0.642469

2B. Comparing > 2 groups means via pairwise comparisons and linear regression

rubric={reasoning:2}

- Use `pairwise.t.test()` to perform pairwise comparisons between groups with corrections for multiple comparisons
- Use `lm()` to perform a linear regression in R to compare the groups (you must adjust for multiple comparisons, do this by performing a Bonferroni correction using `p.adjust()` with `method = "bonferroni"`*)
- Explore `broom::tidy()` to get the results from all the tests above.
- Report the results obtained in a nice data frame or table

Solutions

One-way ANOVA to test our null hypothesis (assess whether the runner's mea speed for all the three footwears are equal):

```r
# formula/model for ANOVA and Linear regression
lm_model <- lm(marathon_ful$mf_s ~ marathon_ful$footwear)

# ANOVA
summary(aov(lm_model))
```

```
##                        Df Sum Sq Mean Sq F value   Pr(>F)    
## marathon_ful$footwear   1    6.8   6.846   19.65 1.04e-05 ***
## Residuals             927  323.0   0.348                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our p-value for the effect of `footwear` is much less than 0.01. We can reject the null hypothesis that the runner's mean speed for all three footwears are equal. To determine which of the groups are different from each other, we can perform a pair-wise comparison between groups (using a Bonferroni correction to control for multiple comparisons):

```r
pairwise_ttest <- tidy(pairwise.t.test(marathon_ful$mf_s, marathon_ful$footwear, p.adjust="bonf"))

print(pairwise_ttest)
```

```
##   group1 group2      p.value
## 1      2      1 0.0001217142
## 2      3      1 0.0384995071
## 4      3      2 0.5134384227
```

The mean speed across all footwears is statistically different, except for footwears 2 and 3. Next, we perform a similar analysis using linear regression:

```r
summary(lm_model)
```

```
## 
## Call:
## lm(formula = marathon_ful$mf_s ~ marathon_ful$footwear)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.38871 -0.40233 -0.00634  0.38893  1.90418 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            3.61118    0.08266  43.687  < 2e-16 ***
## marathon_ful$footwear -0.19793    0.04465  -4.433 1.04e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5903 on 927 degrees of freedom
## Multiple R-squared:  0.02076,    Adjusted R-squared:  0.0197
## F-statistic: 19.65 on 1 and 927 DF,  p-value: 1.041e-05
```

The p-value returned from the linear model comparing a model of 3 footwears to a model with only an intercept is much less than 0.01, and we actually get the same p-value as we get from the ANOVA. This means we can reject the null-hypothesis that the fit of the intercept-only model and our model including `footwear` are equal, and we can conclude that the model with `footwear` provides a better fit than the intercept-only model.

What about the effect of each footwear? To assess which might be different we need to look at the output (p-value) for each footwear compared to the reference. By default, our reference footwear is 1:

```r
marathon_ful <- marathon %>%
  filter(cohort1 == 1) %>%
  select(c(age, bmi, female, footwear, group, injury, mf_d, mf_di, mf_ti,  max, sprint)) %>%
  mutate(mf_s = mf_d/mf_ti, footwear = as.factor(footwear))

lm_model <- lm(marathon_ful$mf_s ~ marathon_ful$footwear)
summary(lm_model)
```

```
##
## Call:
## lm(formula = marathon_ful$mf_s ~ marathon_ful$footwear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38970 -0.40332 -0.00733  0.38989  1.90319
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.41146    0.04186  81.490  < 2e-16 ***
## marathon_ful$footwear2 -0.19514    0.04732  -4.124 4.06e-05 ***
## marathon_ful$footwear3 -0.42149    0.16906  -2.493   0.0128 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5906 on 926 degrees of freedom
## Multiple R-squared:  0.02079,    Adjusted R-squared:  0.01868
## F-statistic: 9.831 on 2 and 926 DF,  p-value: 5.958e-05
```

The linear model suggests that the runner's mean speed of footwear 1 is significantly different than those of footwears 2 and 3 at the 5% level. This result differs from our pairwise comparison, and we have not made all the same comparisons that we did in our pairwise comparison test. Thus to do an equivalent analysis, we might need to repeat the linear regression with different reference groups to make all relevant comparisons, and then adjust the p-values to control for these multiple comparisons:

```r
# footwear 1 as reference:
summary(lm_model)
```

```
##
## Call:
## lm(formula = marathon_ful$mf_s ~ marathon_ful$footwear)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.38970 -0.40332 -0.00733  0.38989  1.90319
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.41146    0.04186  81.490  < 2e-16 ***
## marathon_ful$footwear2 -0.19514    0.04732  -4.124 4.06e-05 ***
## marathon_ful$footwear3 -0.42149    0.16906  -2.493   0.0128 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5906 on 926 degrees of freedom
## Multiple R-squared:  0.02079,    Adjusted R-squared:  0.01868
## F-statistic: 9.831 on 2 and 926 DF,  p-value: 5.958e-05
```

```r
# change the reference level

selectReferenceLMFit <- function(dataset, reference){
   data_ref <- within(dataset, footwear <- relevel(footwear, ref = reference))
   model_fit_ref <- lm(data_ref$mf_s ~ data_ref$footwear)
   lm_result_ref <-  summary(model_fit_ref)
   return(lm_result_ref)
   }

# footwear 2 as reference:
 selectReferenceLMFit(marathon_ful,  "2")
```

```
##
## Call:
## lm(formula = data_ref$mf_s ~ data_ref$footwear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38970 -0.40332 -0.00733  0.38989  1.90319
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.21632    0.02205 145.833  < 2e-16 ***
## data_ref$footwear1  0.19514    0.04732   4.124 4.06e-05 ***
## data_ref$footwear3 -0.22635    0.16527  -1.370    0.171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5906 on 926 degrees of freedom
## Multiple R-squared:  0.02079,    Adjusted R-squared:  0.01868
## F-statistic: 9.831 on 2 and 926 DF,  p-value: 5.958e-05
```

```r
# footwear 3 as reference:
selectReferenceLMFit(marathon_ful,  "3")
```

```
##
## Call:
## lm(formula = data_ref$mf_s ~ data_ref$footwear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.38970 -0.40332 -0.00733  0.38989  1.90319
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.9900     0.1638  18.255   <2e-16 ***
## data_ref$footwear1    0.4215     0.1691   2.493   0.0128 *
## data_ref$footwear2    0.2264     0.1653   1.370   0.1711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5906 on 926 degrees of freedom
## Multiple R-squared:  0.02079,    Adjusted R-squared:  0.01868
## F-statistic: 9.831 on 2 and 926 DF,  p-value: 5.958e-05
```

Similar to the pairwise comparisons, the mean speed across all footwears is statistically different, except for footwears 2 and 3.

Adjusting the linear regression to compare how similar or different the results are:

```r
# footwear 2 as reference:
p.adjust(selectReferenceLMFit(marathon_ful,  "2")$coefficients[,4], method="bonferroni")
```

```
##        (Intercept) data_ref$footwear1 data_ref$footwear3
##       0.0000000000       0.0001217142       0.5134384227
```

```r
# footwear 3 as reference:
p.adjust(selectReferenceLMFit(marathon_ful,  "3")$coefficients[,4], method="bonferroni")
```

```
##        (Intercept) data_ref$footwear1 data_ref$footwear2
##       2.368487e-63       3.849951e-02       5.134384e-01
```

Now the mean speed across all footwears is statistically different.

2C. What can we conclude from applying these different methods to this case?

rubric={reasoning:2}

- Discuss the results you obtained from applying the different methods in Part B to the same question. Do you get the same results or different? Is this what you expected? Why or why not?
- Discuss why you needed to control for multiple comparisons (lean on what you learned in lectures 1 in DSCI_552_stat-inf-1 to answer this question).

Solutions

When we consider a set of hypothesis testing simultaneously, it is always possible that we obtain significance for certain tests purely by chance (we define this as alpha, and typically choose 0.05 for this value). The chances of finding at least one such significant difference increase quite rapidly as the number of tests increases which is why we need to control it.