Homework 1 Solutions

Alexander McLain

1. This question will use data on Orthodontic Measurements on Children. The data are from a study of dental growth measurements of the distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure. Measurements were obtained on 11 girls and 16 boys at ages 8, 10, 12, and 14. Scientific goals of the study are to:

- determine whether distances over time are larger for boys than for girls, and
- whether the rate of change of distance over time is similar for boys and girls.

The data is available on the website in the file `dental.csv`, which contains the data and information about the study and variables.

First, I'll load in the data set, create a "long'' version and peak at the variables. Notice that for the `pitot_longer` command I made some changes versus what we had in the first example.

```
library(tidyverse)
setwd("~/Library/CloudStorage/OneDrive-UniversityofSouthCarolina/Teaching/755_Spring/Homework")
wide_dental <- read.csv("dental.csv",header = TRUE, na.strings = "",
                        stringsAsFactors = FALSE)
head(wide_dental)
```

ID	Gender	Dist8	Dist10	Dist12	Dist14
1	F	21.0	20.0	21.5	23.0
2	F	21.0	21.5	24.0	25.5
3	F	20.5	24.0	24.5	26.0
4	F	23.5	24.5	25.0	26.5
5	F	21.5	23.0	22.5	23.5
6	F	20.0	21.0	21.0	22.5

Now, we'll go from wide to long.

```
long_dental <- pivot_longer(wide_dental, cols = starts_with("Dist"), names_to = "Age",
                            names_prefix = "Dist", values_to = "Dist",
                            values_drop_na = TRUE)
str(long_dental)
```

```
## tibble [108 x 4] (S3: tbl_df/tbl/data.frame)
##  $ ID    : int [1:108] 1 1 1 1 2 2 2 2 3 3 ...
##  $ Gender: chr [1:108] "F" "F" "F" "F" ...
##  $ Age   : chr [1:108] "8" "10" "12" "14" ...
##  $ Dist  : num [1:108] 21 20 21.5 23 21 21.5 24 25.5 20.5 24 ...
```

Notice that `Age` has *chr* in it's heading. This indicates that `R` is treating it as a charater variable (not a numeric variable). I would rather have `Age` as a numeric variable for plotting purposes and will change it to one using the `as.numeric` function. **(Note: if you skip this step the plots will go from 10, 12, 14, to 8.)**

```
long_dental <- long_dental %>% mutate( Age = as.numeric(Age) )
str(long_dental)
```

```
## tibble [108 x 4] (S3: tbl_df/tbl/data.frame)
##  $ ID    : int [1:108] 1 1 1 1 2 2 2 2 3 3 ...
##  $ Gender: chr [1:108] "F" "F" "F" "F" ...
##  $ Age   : num [1:108] 8 10 12 14 8 10 12 14 8 10 ...
##  $ Dist  : num [1:108] 21 20 21.5 23 21 21.5 24 25.5 20.5 24 ...
```

OK. Now we're good and ready to use this dataset for analysis.

a. **(15 Points) Read the data into R and calculate sample means, standard deviations and variances of the distance measurements at each occasion.**

There are many ways to get these statistics. I'm going to show you a way I think is simple using the `apply` function. This function is specifically made to get column or row statistics. This function takes 3 arguements.

```
args(apply)
```

```
## function (X, MARGIN, FUN, ..., simplify = TRUE)
## NULL
```

The first argument is the data, for the second you put '1' if you want row statistics (i.e., the average per individual) and '2' if you want column statistics (what we want), and for the third you put what statistics you want. Here we want the mean (`mean`), standard deviation (`sd`) and variance (`var`).

Also, i'm going to remove the first two rows since we don't want the mean ID or gender.

```
means <- apply(wide_dental[ , - c(1,2)], 2, mean)
sds   <- apply(wide_dental[ , - c(1,2)], 2, sd)
vars  <- apply(wide_dental[ , - c(1,2)], 2, var)
```

Now that they been calculated I'm going to display them together with a data frame (this is not required).

```
data.frame(means = means, sds = sds, vars = vars)
```

	means	sds	vars
Dist8	22.18519	2.434322	5.925926
Dist10	23.16667	2.157278	4.653846
Dist12	24.64815	2.817578	7.938746
Dist14	26.09259	2.766687	7.654558

b. **(15 Points) Construct a spaghetti plot for 10 subjects (5 girls, 5 boys). Make the plot so that the point symbols are different for girls and boys.**

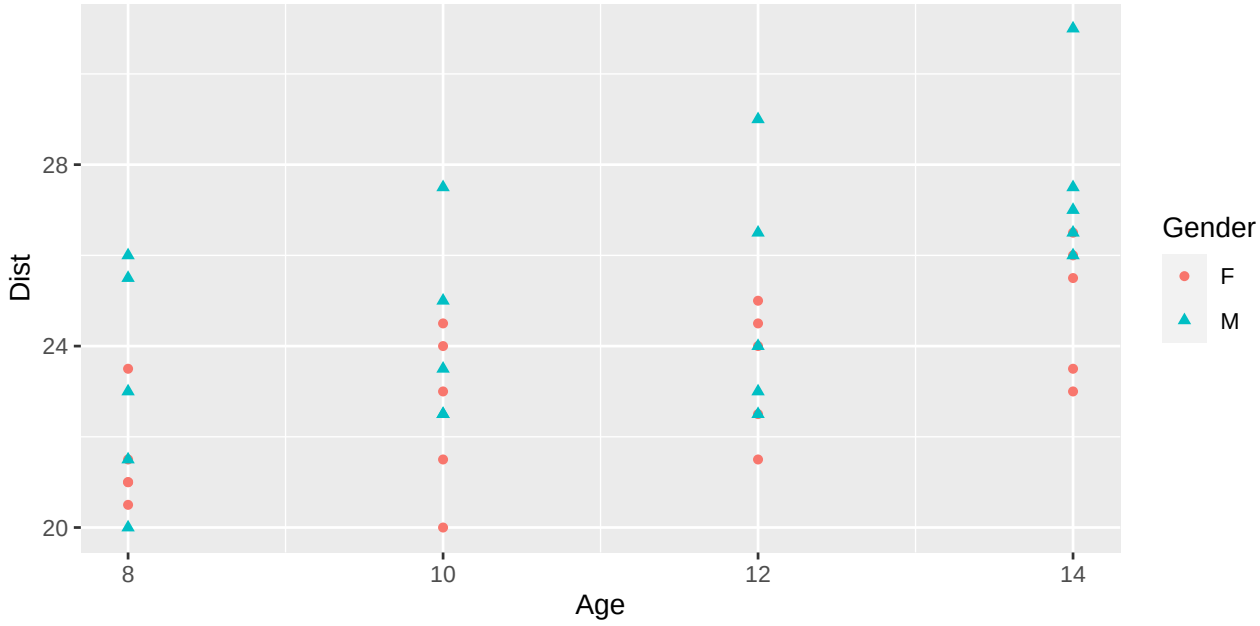
There are many ways to get your data set down to 5 girls and 5 boys. A simple way would be to modify the .csv file in excel and re-read the dataset in. Here i'm going to `filter` the data by the ID's that I want. The first 5 ID's for the first are 1, 2, 3, 4, and 5 (which in R can be represented as `1:5`) and the first 5 ID's for the boys are 12, 13, . . . , 16 (or `12:16`).

So I want to filter the data so that my ID's are *in* `c(1:5, 12:16)` we can do that with the `%in%` command. Putting this together we have:

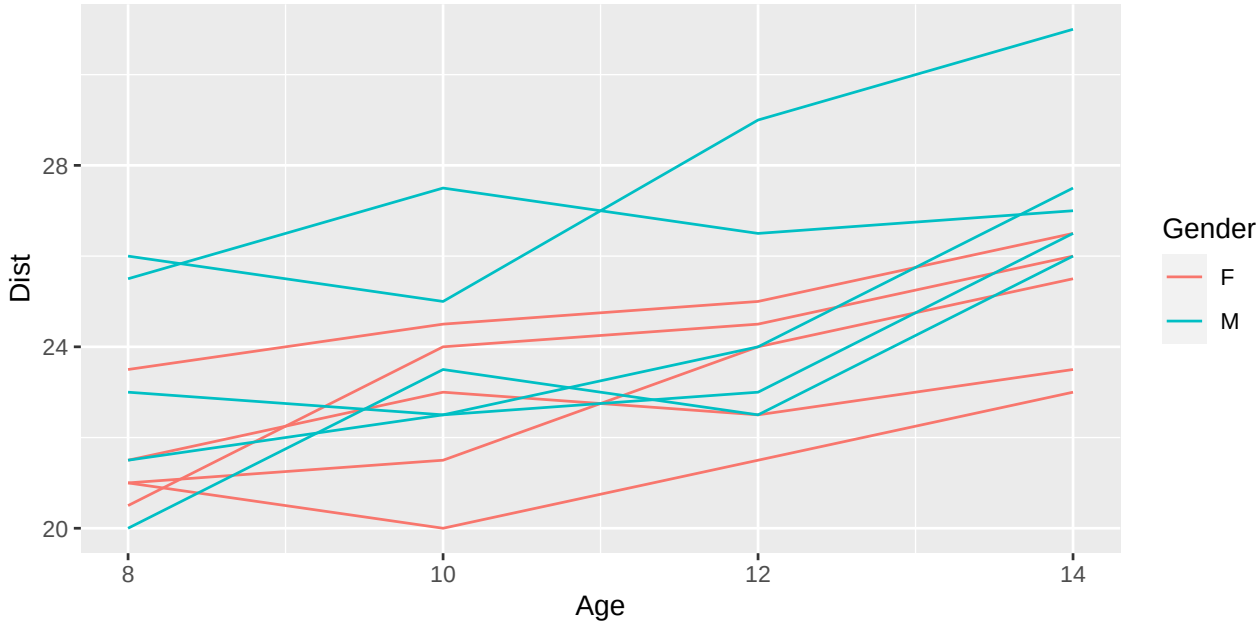
```
long_dental_sm <- long_dental %>% filter( ID %in% c(1:5, 12:16))
```

Now we can do the point plot like we did in class. Here i'm using different point symbols and colors by gender.

```
p <- ggplot(data = long_dental_sm, aes(x = Age, y = Dist, group = ID))
p + geom_point(aes(shape = Gender, color = Gender))
```




```
p + geom_line(aes(color = Gender))
```



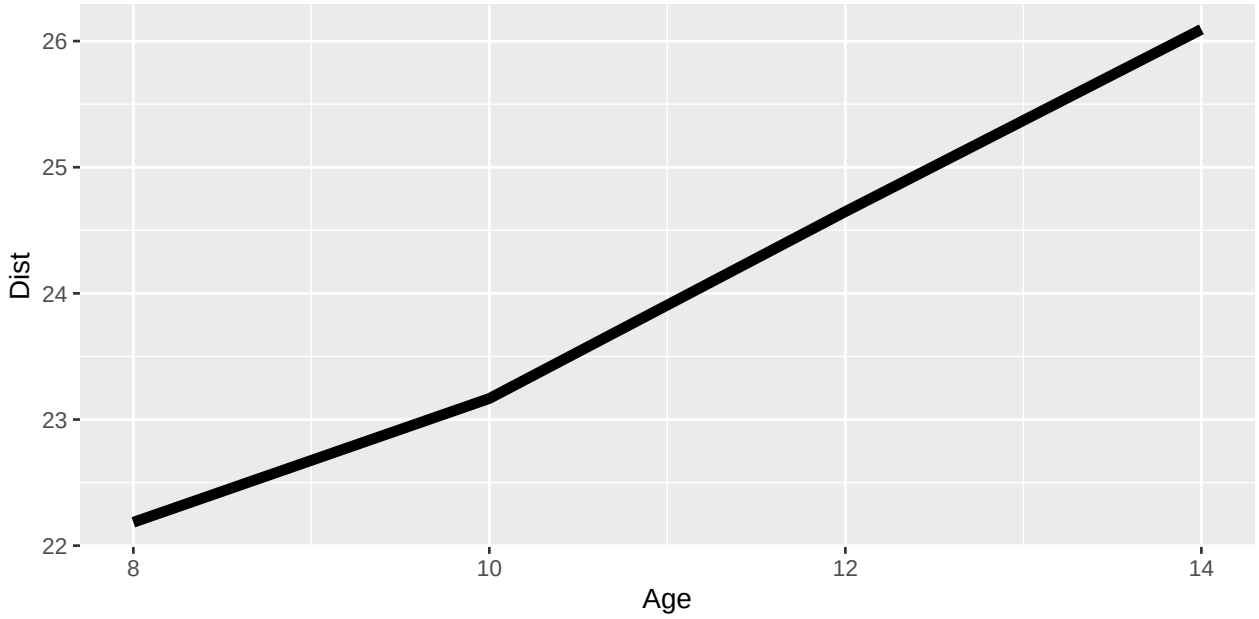
c. **(15 Points) Construct a time plot of the mean measurement versus time (in years). Initially do this for all subjects, then construct a plot that contains two lines: one describing the mean measurement for boys and one the mean measurement for girls. Make sure it is possible to differentiate the lines using different colors or line types.**

This example follows from the one we did in class. Note we'll go back to using the full dataset (not the small version).

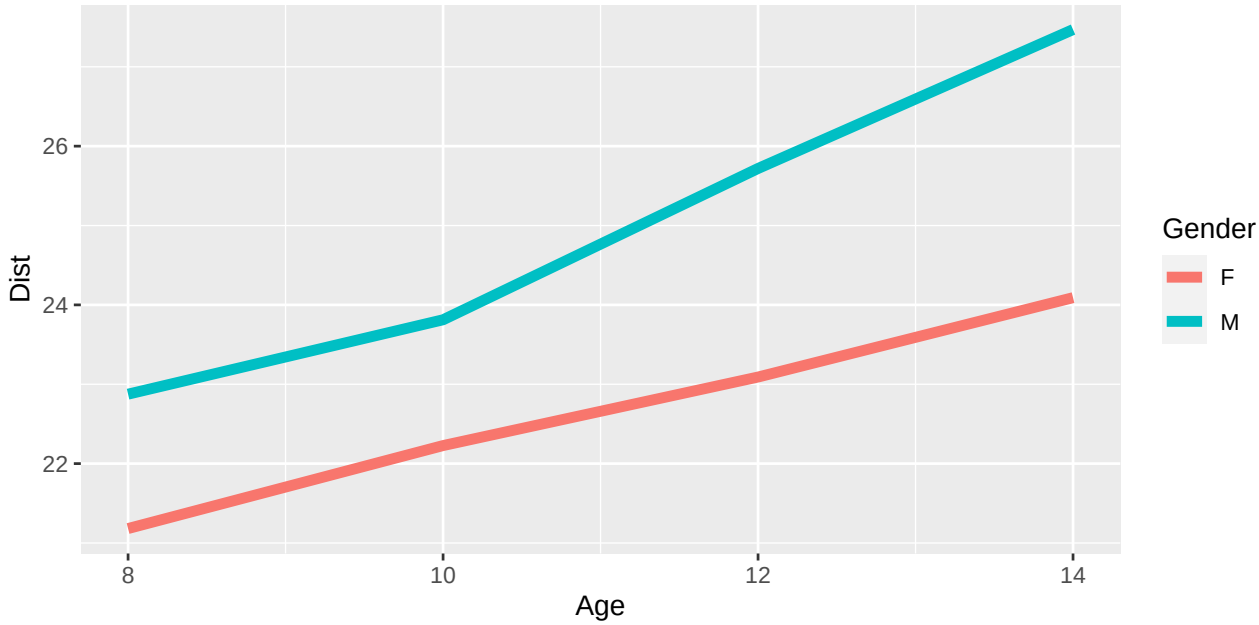
```
p <- ggplot(data = long_dental, aes(x = Age, y = Dist, group = ID))
p + stat_summary(aes(group = 1), geom = "line",
```

```
                  fun = mean, size = 2)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
p + stat_summary(aes(group = Gender, color = Gender), geom = "line",
                  fun = mean, size = 2)
```



d. **(15 Points) Calculate the 4 x 4 covariance and correlation matrices for the four repeated measures. Does correlation appear to vary by the time difference between measurements?**

This will be done similarly to the example in class.

```r
cov(wide_dental[ , -c(1,2)])
```

	Dist8	Dist10	Dist12	Dist14
Dist8	5.925926	3.285256	4.875356	4.039886
Dist10	3.285256	4.653846	3.858974	4.532051
Dist12	4.875356	3.858974	7.938746	6.197293
Dist14	4.039886	4.532051	6.197293	7.654558

```r
cor(wide_dental[ , -c(1,2)])
```

	Dist8	Dist10	Dist12	Dist14
Dist8	1.0000000	0.6255833	0.7108079	0.5998338
Dist10	0.6255833	1.0000000	0.6348775	0.7593268
Dist12	0.7108079	0.6348775	1.0000000	0.7949980
Dist14	0.5998338	0.7593268	0.7949980	1.0000000

There does not appear to be much difference in the correlations that are 1 measurement apart (0.62, 0.63, 0.79) and those that are 2 measurements apart (0.71, 0.76). The correlation between the measurments that are 3 apart (0.59) is the lowest of the group, but not by much. Overall, it's hard to see much of a pattern here.

e. **(15 Points) Calculate the 4 x 4 covariance and correlation matrices for the four repeated measures separately for boys and girls. Comment on the differences in the variance and correlation between boys and girls.**

Do do this we'll first create separate data sets for boys and girls. Then estimate the covariance and correlation matrix for each.

```r
wide_dental_G <- wide_dental %>% filter( Gender == "F")
wide_dental_B <- wide_dental %>% filter( Gender == "M")
# Covariance and correlation for girls.
cov(wide_dental_G[ , -c(1,2)])
```

	Dist8	Dist10	Dist12	Dist14
Dist8	4.513636	3.354545	4.331818	4.356818
Dist10	3.354545	3.618182	4.027273	4.077273
Dist12	4.331818	4.027273	5.590909	5.465909
Dist14	4.356818	4.077273	5.465909	5.940909

```r
cor(wide_dental_G[ , -c(1,2)])
```

	Dist8	Dist10	Dist12	Dist14
Dist8	1.0000000	0.8300900	0.8623146	0.8413558
Dist10	0.8300900	1.0000000	0.8954156	0.8794236
Dist12	0.8623146	0.8954156	1.0000000	0.9484070
Dist14	0.8413558	0.8794236	0.9484070	1.0000000

```r
# Covariance and correlation for boys.
cov(wide_dental_B[ , -c(1,2)])
```

	Dist8	Dist10	Dist12	Dist14
Dist8	6.016667	2.291667	3.629167	1.612500
Dist10	2.291667	4.562500	2.193750	2.810417
Dist12	3.629167	2.193750	7.032292	3.240625
Dist14	1.612500	2.810417	3.240625	4.348958

```r
cor(wide_dental_B[ , -c(1,2)])
```

	Dist8	Dist10	Dist12	Dist14
Dist8	1.0000000	0.4373932	0.5579310	0.3152311
Dist10	0.4373932	1.0000000	0.3872909	0.6309234
Dist12	0.5579310	0.3872909	1.0000000	0.5859866
Dist14	0.3152311	0.6309234	0.5859866	1.0000000

The variance of the boys measurements are larger for all but the last measurement. The correlation between measurments are markedly higher for the girls versus the boys.

f. **(15 Points) Recall the 4 correlations "truths" from the second lecture. Describe on what these truths mean in the context of this example. In your description, try to be more specific than restating each "truth" in context. For example, don't use the word 'correlation' in your description, think about what the "truth" tells you about the data if taken as fact.**

The 4 correlation "truths" are:

i. the correlations are positive,
ii. the correlations often decrease with increasing time separation,
iii. the correlations between repeated measures rarely ever approach zero, and
iv. the correlation between a pair of repeated measures taken very closely rarely approaches one.

In this example, these truths will mean:

i. If a childs dental growth at one measurement is higher (lower) than average we would expect their other measurments to be higher (lower) than average.
ii. If a childs dental growth is higher than average at the first measurement we are confident the next measurement will be higher than average, slight less confident for the measurement after that and even less confident for the measurement after that.
iii. If a childs dental growth is higher than average at the first measurement a subsequent measurment (no matter how long after it is taken) is more likely to be higher than average.
iv. If a childs dental growth at one measurement is higher than average, than the probability that a different measurement is higher than average is less than one no matter how closely the measurments are taken.

g. **(10 Points) State three potential sources of variability in the data in context of the problem. Using the time plot you constructed in (b) give your hypothesized ordering of which source of variability is the largest.**

The 3 sources of variability are:

i. between individual variation,
ii. within individual variation, and
iii. measurement error.

Given the time plot I would say there is similar between individual and within individual variation in the data. There also appears to be a lot of measurment error judging by the one measurment that has a big increase followed by a big decrease (obviously we wouldn't expect dental measurements to decrease).