Multiple Statistical Analyses

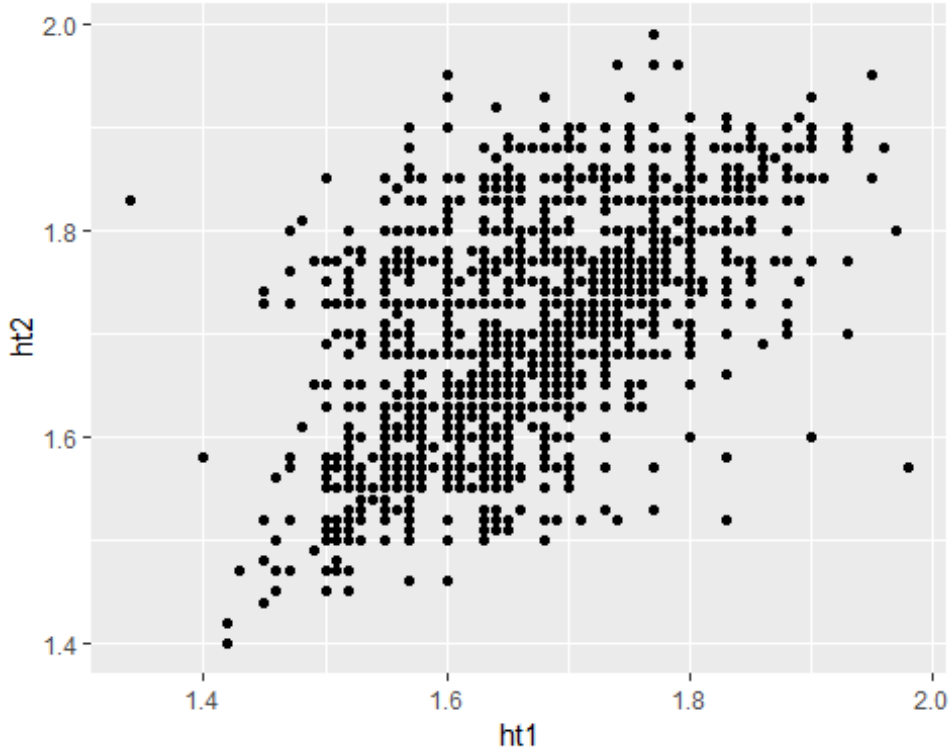
Yogindra Raghav

November 16, 2018

Using ggplot2, create scatter plots to visually investigate the relation between ht1 and ht2. The scatter plot gives some impression on the answer of the question, “yes”. Why?

```
twinData %>% ggplot(aes(ht1, ht2)) + geom_point()

## Warning: Removed 141 rows containing missing values (geom_point).
```

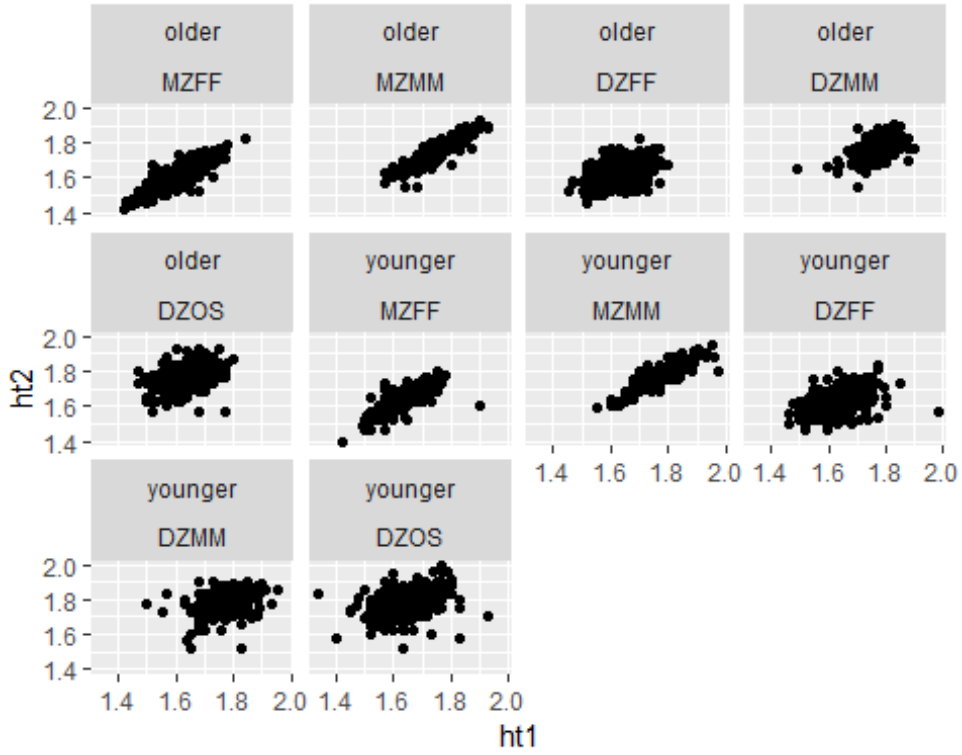


Most of the points on this scatter plot seem to fall on a linear regression line. If we were given the equation of the regression line, we should be able to predict the height of the other twin.

But, will the answer be the same if we use more information? Consider two more variables cohort, zygosity. Add more layers or use facet to include the information contained in these two variables.

```
twinData %>% ggplot(mapping = aes(ht1, ht2)) + geom_point()+ facet_wrap( coho
rt ~zygosity)

## Warning: Removed 141 rows containing missing values (geom_point).
```



The linear trend mentioned earlier is still present even when separating the data into these facets.

Sort the result of above computation from the largest estimate of correlation coefficient to the smallest.

```
library(broom)
twinData %>% group_by(cohort,zygosity) %>%  do(tidy( cor.test(ht1 ~ ht2, alte
rnative = "greater" , . ))) %>% arrange(desc(estimate))
```

```
## # A tibble: 10 x 10
## # Groups:   cohort, zygosity [10]
##    cohort zygosity estimate statistic   p.value parameter conf.low
##    <chr>  <fct>       <dbl>     <dbl>     <dbl>     <int>    <dbl>
##  1 older  MZMM        0.907     36.4  1.09e-109       286    0.888
##  2 young~ MZMM        0.883     30.0  3.36e- 86       256    0.858
##  3 young~ MZFF        0.877     42.7  8.60e-177       547    0.860
##  4 older  MZFF        0.859     42.6  1.66e-189       642    0.841
##  5 older  DZMM        0.510      6.97 5.84e- 11       138    0.399
##  6 older  DZFF        0.456     10.1  1.06e- 21       387    0.388
##  7 young~ DZFF        0.440      9.02 7.36e- 18       339    0.365
##  8 young~ DZOS        0.428     10.4  2.19e- 23       484    0.365
##  9 older  DZOS        0.381      8.00 7.65e- 15       378    0.306
## 10 young~ DZMM        0.350      5.15 3.32e-  7       190    0.241
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

Create a new variable to indicate whether the correlation coefficient between ht1 and ht2 in the particular subgroup is greater 0.5, with 95 percent confidence. Save the resulting data frame by the name sig_twin_cor.

```
Sig_twin_cor <- twinData %>% group_by(cohort,zygosity) %>%  do(tidy( cor.test
(ht1 ~ ht2, alternative = "greater" , . ))) %>% arrange(desc(estimate))


sig_twin_cor$corr_sig[twin_cor$conf.low>=0.5] <- 1
```

```
## Warning: Unknown or uninitialised column: 'corr_sig'.
```

```
Sig_twin_cor$corr_sig[twin_cor$conf.low<=0.5] <- 0
```

List the only the combinations of cohort and zygosity where the twins' heights are significantly similar. Here the similarity is defined by the test result, evaluated in #6.

```
sig_twin_cor %>% filter(corr_sig == 1)
```

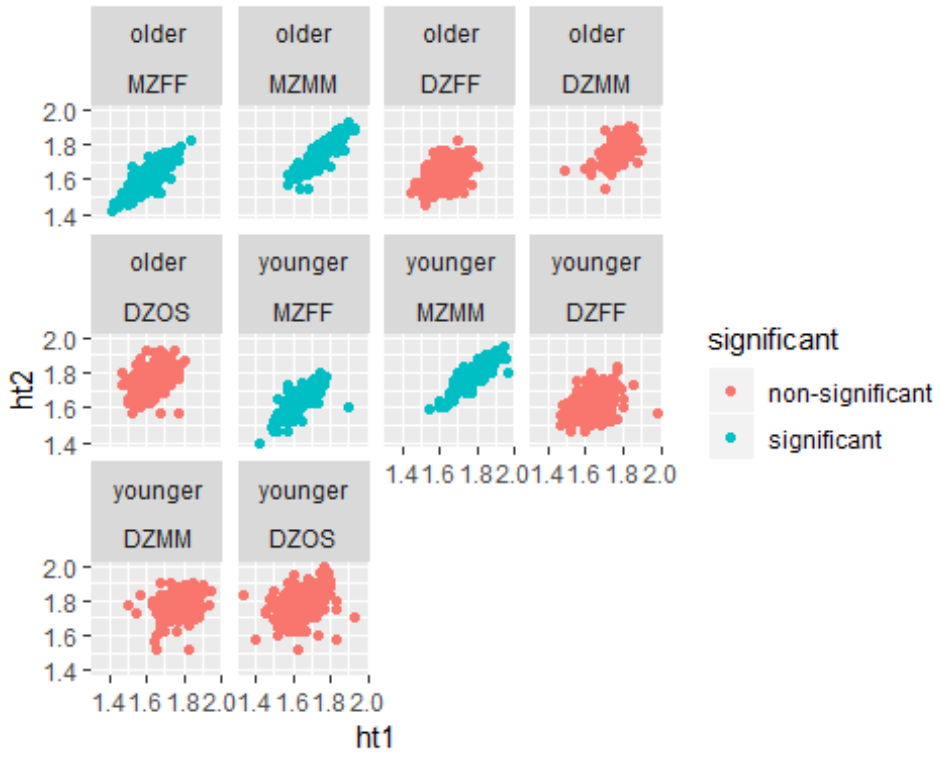
```
## # A tibble: 4 x 11
## # Groups:   cohort, zygosity [4]
##   cohort zygosity estimate statistic   p.value parameter conf.low conf.hig
h
##   <chr>  <fct>       <dbl>     <dbl>     <dbl>     <int>    <dbl>     <dbl
>
## 1 older  MZMM        0.907      36.4 1.09e-109       286    0.888
1
## 2 young~ MZMM        0.883      30.0 3.36e- 86       256    0.858
1
## 3 young~ MZFF        0.877      42.7 8.60e-177       547    0.860
1
## 4 older  MZFF        0.859      42.6 1.66e-189       642    0.841
1
## # ... with 3 more variables: method <chr>, alternative <chr>,
## #   corr_sig <dbl>
```

Repeat exercise #3. This time, use the variables cohort, zygosity to facet, and use different colors to indicate the subgroups for which the heights are significantly similar. Comment on your finding.

```
twin_2_data = twinData %>% mutate(significant = ifelse(zygosity == "MZMM" | z
ygosity == "MZFF", "significant", "non-significant" ))


twin_2_data %>% ggplot(mapping = aes(ht1, ht2, color = significant )) + geom_
point()+ facet_wrap(cohort ~ zygosity)
```

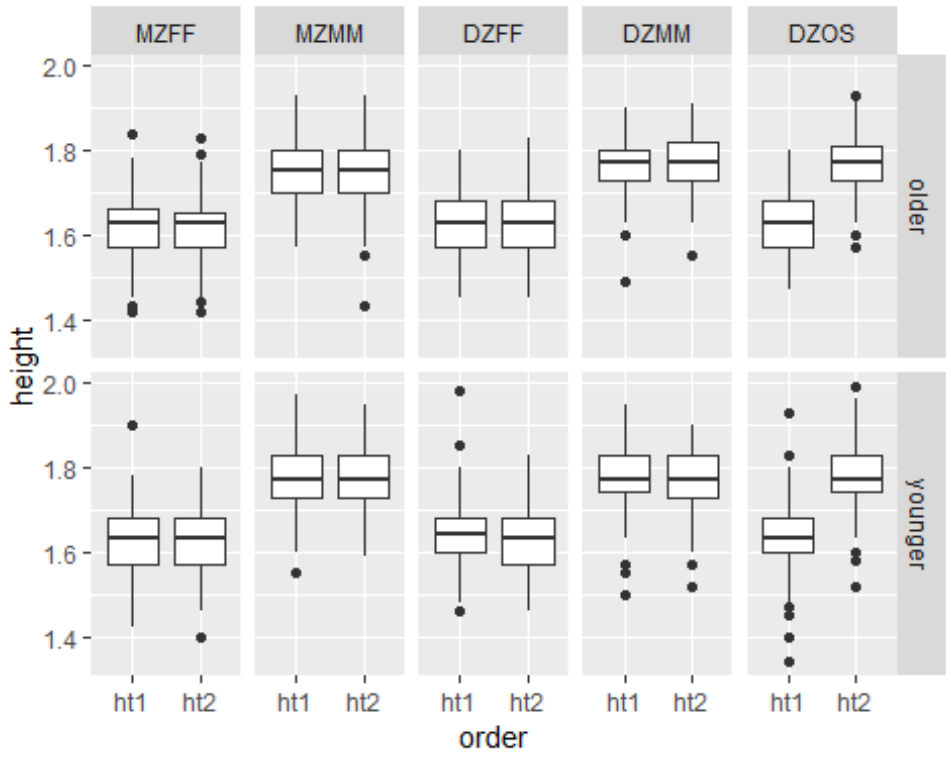
```
## Warning: Removed 141 rows containing missing values (geom_point).
```



Recreate the following graphic. This involves transforming twinData into a narrow form using gather(). You might want to take a look at Lecture 5 slides for boxplots.

```
twinData %>% gather('ht1', 'ht2', key = "order", value = "height") %>% ggplot
(aes(order))+ geom_boxplot(aes(y = height))+ facet_grid(cohort~zygosity)

## Warning: Removed 150 rows containing non-finite values (stat_boxplot).
```



Inspect the data graphic. Is there any need to adjust the hypothesis (posed in Question #2)?

No because the medians of all zygosities except for DZOS are almost the same. We might need to make a new hypothesis for the DZOS zygosity.

Recreate the above graphic with different colors indicating the results of t-tests (based on p-value).

```
twinData %>% select(cohort,zygosity,ht1,ht2) %>% group_by(cohort,zygosity) %>
% do(tidy(t.test(.$ht1, .$ht2, paired = TRUE)))
```

```
## # A tibble: 10 x 10
## # Groups:   cohort, zygosity [10]
##    cohort zygosity estimate statistic   p.value parameter conf.low
##    <chr>  <fct>       <dbl>     <dbl>     <dbl>     <dbl>    <dbl>
##  1 older  MZFF      1.27e-3     0.953 3.41e-  1       643 -1.35e-3
```

```
##  2 older  MZMM     -2.39e-4    -0.131 8.95e-  1       287 -3.82e-3
##  3 older  DZFF      3.14e-3     0.916 3.60e-  1       388 -3.60e-3
##  4 older  DZMM     -6.35e-3    -1.20  2.31e-  1       139 -1.68e-2
##  5 older  DZOS     -1.41e-1   -40.5   9.99e-140       379 -1.48e-1
##  6 young~ MZFF      1.80e-4     0.128 8.98e-  1       548 -2.58e-3
##  7 young~ MZMM      1.28e-3     0.635 5.26e-  1       257 -2.70e-3
##  8 young~ DZFF      7.61e-3     1.94  5.30e-  2       340 -9.86e-5
##  9 young~ DZMM      2.13e-3     0.376 7.07e-  1       191 -9.03e-3
## 10 young~ DZOS     -1.43e-1   -43.2   3.16e-168       485 -1.49e-1
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

```
twinData %>% mutate(is.DZOS = ifelse(zygosity == "DZOS", "DZOS", "non-DZOS" )
) %>% gather('ht1', 'ht2', key = "order", value = "height") %>% ggplot(aes(or
der, color = is.DZOS))+ geom_boxplot(aes(y = height))+ facet_grid(cohort~zygo
sity)
```

```
## Warning: Removed 150 rows containing non-finite values (stat_boxplot).
```