HW6

Laura Eldridge

4/12/2020

Download Data

```
pete <- read.csv("C:\\Users\\Laura\\Desktop\\Stats 488\\Homework\\ill_school_data.csv", header = T)
```

Describe the Data

```
summary(pete)
```

```
##  Country   Region      DataYear      ClassGrade       Gender
##  USA:500   IL:500   Min.   :2012   Min.   : 9.00         :  5
##                     1st Qu.:2015   1st Qu.:10.00   Female:274
##                     Median :2017   Median :12.00   Male  :221
##                     Mean   :2017   Mean   :10.88
##                     3rd Qu.:2018   3rd Qu.:12.00
##                     Max.   :2019   Max.   :12.00
##
##     Ageyears              Handed      Height_cm   Footlength_cm
##  Min.   :12.00                : 12   165    : 30   24     : 57
##  1st Qu.:15.00   Ambidextrous: 19          : 29   23     : 54
##  Median :17.00   Left-Handed : 47   170    : 29   25     : 51
##  Mean   :16.36   Right-Handed:422   160    : 25   26     : 44
##  3rd Qu.:17.00                      163    : 19          : 34
##  Max.   :99.00                      162    : 14   22     : 32
##  NA's   :4                          (Other):354   (Other):228
##    Armspan_cm  Languages_spoken                Travel_to_School
##         : 49   Min.   : 0.00                           : 21
##  160    : 30   1st Qu.: 1.00    Bicycle                :  5
##  170    : 18   Median : 2.00    Boat                   :  6
##  162    : 15   Mean   : 1.69    Bus                    : 69
##  180    : 15   3rd Qu.: 2.00    Car                    :346
##  163    : 13   Max.   :13.00    Rail (Train/Tram/Subway):  7
##  (Other):360   NA's   :22       Walk                   : 46
##  Travel_time_to_School Reaction_time Score_in_memory_game
##  15     : 79                  : 32          : 35
##  10     : 70           0.3    :  6   40     : 33
##  20     : 55           0.343  :  5   44     : 24
##  5      : 36           0.36   :  5   38     : 23
##  7      : 32           0.43   :  5   45     : 22
##  30     : 26           0.446  :  5   46     : 22
##  (Other):202           (Other):442   (Other):341
##  Favourite_physical_activity Importance_reducing_pollution
##  Other     : 74              1000   :104
##  Soccer    : 71                     : 38
##  Basketball: 41              500    : 33
##  Swimming  : 37              900    : 32
##            : 28              800    : 27
##  Tennis    : 28              0      : 21
##  (Other)   :221              (Other):245
##  Importance_recycling_rubbish Importance_conserving_water
##  Min.   :   0.0               Min.   :   0.0
##  1st Qu.: 439.0               1st Qu.: 400.5
##  Median : 600.0               Median : 652.0
##  Mean   : 631.3               Mean   : 628.9
##  3rd Qu.: 850.0               3rd Qu.: 925.0
##  Max.   :9050.0               Max.   :5000.0
##  NA's   :47                   NA's   :49
##  Importance_saving_energy Importance_owning_computer
##  1000   : 76              1000   : 57
##         : 50                     : 50
##  500    : 40              500    : 33
##  800    : 27              0      : 27
```

```
##  600    : 25              100    : 24
##  700    : 25              800    : 22
##  (Other):257              (Other):287
##  Importance_Internet_access Left_Footlength_cm      Longer_foot
##  1000   :138                       : 63                   : 59
##         : 51                24     : 59        Left foot  : 96
##  500    : 26                25     : 50        Right foot : 88
##  700    : 25                23     : 47        Same length:257
##  800    : 23                26     : 42
##  900    : 19                22     : 38
##  (Other):218                (Other):201
##  Index_Fingerlength_mm Ring_Fingerlength_mm  Longer_Finger_Lefthand
##         : 73                   : 74                       : 58
##  70     : 48           70     : 49          Index finger:137
##  80     : 40           80     : 45          Ring finger :214
##  75     : 32           7      : 23          Same length : 91
##  90     : 21           75     : 19
##  65     : 20           90     : 19
##  (Other):266           (Other):271
##     Birth_month  Favorite_Season Allergies Vegetarian
##           : 50         : 50         : 52      : 52
##  September: 49   Fall  :178      No :305   No :423
##  June     : 47   Spring: 57      Yes:143   Yes: 25
##  May      : 46   Summer:168
##  October  : 43   Winter: 47
##  February : 36
##  (Other)  :229
##             Favorite_Food                     Beverage
##  Meat              :101   Water                   :285
##  Pizza/Pasta       : 96                           : 52
##  No favorite       : 55   Juice                   : 44
##                    : 51   Milk                    : 28
##  Fruit             : 43   Soft drink (caffeinated): 28
##  Rice/Noodle dishes: 41   Tea                     : 17
##  (Other)           :113   (Other)                 : 46
##                Favorite_School_Subject Sleep_Hours_Schoolnight
##  Mathematics and statistics: 82        7      :125
##  English                   : 70        6      :108
##  History                   : 55        8      :102
##  Science                   : 55                : 49
##                            : 53        5      : 37
##  Physical education        : 46        9      : 27
##  (Other)                   :139        (Other): 52
##  Sleep_Hours_Non_Schoolnight Home_Occupants
##  10     :102                  4      :161
##  9      : 96                  5      :103
##  8      : 81                  3      : 87
##         : 52                  6      : 52
##  7      : 27                         : 51
##  11     : 24                  2      : 22
##  (Other):118                  (Other): 24
##                   Home_Internet_Access
##                             : 55
##  No internet connection     :  6
```

```
##  Yes - broadband connection:270
##  Yes - dial-up connection  : 14
##  Yes - other               :155
##
##
##                                                 Communication_With_Friends
##                                                                    : 61
##  Cell phone                                                        : 35
##  In person                                                         : 74
##  Internet chat or instant messaging                                : 31
##  Myspace, Facebook, other social networking sites, or blog: 84
##  Other                                                             : 11
##  Text messaging                                                    :204
##  Text_Messages_Sent_Yesterday Text_Messages_Received_Yesterday
##         : 54                          : 56
##  50     : 51                  50     : 42
##  100    : 36                  100    : 35
##  20     : 34                  20     : 24
##  30     : 29                  30     : 20
##  10     : 24                  10     : 19
##  (Other):272                  (Other):304
##  Hanging_Out_With_Friends_Hours Talking_On_Phone_Hours
##         : 58                     1      :124
##  5      : 44                     0      : 94
##  10     : 40                             : 70
##  3      : 35                     2      : 49
##  6      : 29                     3      : 30
##  2      : 28                     0.5    : 24
##  (Other):266                     (Other):109
##  Doing_Homework_Hours Doing_Things_With_Family_Hours
##         : 69                  : 71
##  10     : 44          10     : 50
##  5      : 39          2      : 50
##  2      : 38          5      : 46
##  1      : 35          3      : 42
##  4      : 29          1      : 32
##  (Other):246          (Other):209
##  Outdoor_Activities_Hours Video_Games_Hours Social_Websites_Hours
##         : 75              0      :153              : 79
##  0      : 52                     : 78       10     : 40
##  10     : 50              1      : 58       5      : 38
##  5      : 37              5      : 31       3      : 32
##  2      : 31              2      : 30       1      : 30
##  12     : 28              4      : 20       20     : 30
##  (Other):227              (Other):130       (Other):251
##  Texting_Messaging_Hours Computer_Use_Hours Watching_TV_Hours
##         : 80                     : 79               : 81
##  1      : 62             10     : 54        0      : 63
##  2      : 45             2      : 34        2      : 40
##  5      : 30             5      : 29        3      : 39
##  10     : 29             20     : 27        1      : 38
##  3      : 27             3      : 25        5      : 38
##  (Other):227             (Other):252        (Other):201
##  Paid_Work_Hours Work_At_Home_Hours  Schoolwork_Pressure
```

```
##  0      :248           : 80                  : 76
##         : 79     2     : 80        A lot     :159
##  10     : 19     1     : 77        None      : 13
##  15     : 16     3     : 52        Some      :199
##  5      : 16     0     : 47        Very little: 53
##  20     : 13     5     : 39
##  (Other):109     (Other):125
##           Planned_Education_Level       Favorite_Music
##                         : 75        Rap/Hip hop  :139
##  Graduate degree        :312        Pop          : 87
##  High school            : 13                     : 78
##  Less than high school:  5          Other        : 77
##  Other                  : 28        Country      : 35
##  Some college           : 17        Rock and roll: 18
##  Undergraduate degree : 50          (Other)      : 66
##           Superpower  Preferred_Status           Role_Model_Type
##                : 81          : 75      Relative          :148
##  Fly           : 88   Famous : 22                        : 76
##  Freeze time   :112   Happy  :238      Sports person     : 51
##  Invisibility  : 81   Healthy: 72      Other             : 47
##  Super strength: 20   Rich   : 93      Musician or singer: 41
##  Telepathy     :118                    Friend            : 29
##                                        (Other)           :108
##                     Charity_Donation
##  Health                        :135
##                                : 76
##  International aid             : 67
##  Wildlife, animals             : 60
##  Education/Youth development: 55
##  Environment                   : 44
##  (Other)                       : 63
```

This is a summary of census data collected from 500 students in Illinois. They collected both quantifiable information on the student's statistics of Age, Grade, their dominant hand, height, foot length and personal questions, such as their preferences, opinions and lifestyles.

This data has some parts that seen impossible, it's only high school students, yet the ages span 12-99 years, indicating there is some incorrect data. Also, the data is collected from many different years, giving it a lot of variety. Every section except for data year has missing or blank data.

Testing for Independance

```
attach(pete)
george <- table(Handed,Favorite_Season)
george
```

```
##               Favorite_Season
## Handed             Fall Spring Summer Winter
##                  11    0      0      1      0
##   Ambidextrous    3    7      2      4      3
##   Left-Handed     3   14      8     17      5
##   Right-Handed   33  157     47    146     39
```

```
chisq.test(george, correct = F)
```

```
## Warning in chisq.test(george, correct = F): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  george
## X-squared = 96.46, df = 12, p-value = 2.741e-15
```

If we define alpha as 0.05, we reject the null hypothesis that what hand is dominant is idenpendant of the student's favorite season. We conclude there is some dependence.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
frank <-  read.csv("C:\\Users\\Laura\\Desktop\\Stats 488\\Homework\\ill_school_data.csv", header
= T, na.strings = c("", "NA"))
frank <- frank%>% na.omit()
attach(frank)
```

```
## The following objects are masked from pete:
##
##     Ageyears, Allergies, Armspan_cm, Beverage, Birth_month,
##     Charity_Donation, ClassGrade, Communication_With_Friends,
##     Computer_Use_Hours, Country, DataYear, Doing_Homework_Hours,
##     Doing_Things_With_Family_Hours, Favorite_Food, Favorite_Music,
##     Favorite_School_Subject, Favorite_Season,
##     Favourite_physical_activity, Footlength_cm, Gender, Handed,
##     Hanging_Out_With_Friends_Hours, Height_cm,
##     Home_Internet_Access, Home_Occupants,
##     Importance_conserving_water, Importance_Internet_access,
##     Importance_owning_computer, Importance_recycling_rubbish,
##     Importance_reducing_pollution, Importance_saving_energy,
##     Index_Fingerlength_mm, Languages_spoken, Left_Footlength_cm,
##     Longer_Finger_Lefthand, Longer_foot, Outdoor_Activities_Hours,
##     Paid_Work_Hours, Planned_Education_Level, Preferred_Status,
##     Reaction_time, Region, Ring_Fingerlength_mm, Role_Model_Type,
##     Schoolwork_Pressure, Score_in_memory_game,
##     Sleep_Hours_Non_Schoolnight, Sleep_Hours_Schoolnight,
##     Social_Websites_Hours, Superpower, Talking_On_Phone_Hours,
##     Text_Messages_Received_Yesterday,
##     Text_Messages_Sent_Yesterday, Texting_Messaging_Hours,
##     Travel_time_to_School, Travel_to_School, Vegetarian,
##     Video_Games_Hours, Watching_TV_Hours, Work_At_Home_Hours
```

```
joe <- table(Handed,Favorite_Season)
joe
```

```
##               Favorite_Season
## Handed         Fall Spring Summer Winter
##   Ambidextrous    4      1      3      2
##   Left-Handed    10      5     11      5
##   Right-Handed  116     35    107     25
```

```
chisq.test(joe)
```

```
## Warning in chisq.test(joe): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  joe
## X-squared = 3.6665, df = 6, p-value = 0.7217
```

When the missing data is omited, we fail to reject the null hypothesis that the two variables are independant. Which makes much more sense, because in the real world, there is no possible way these things are connected.

Here the missing data can and should be omitted, because otherwise it clouds the results and gives us an errorous conclusion.

I used the Chi-Squared test because it is the most appropriate for testing independance for multi-leveled variables. Fisher's only works on two by two tables and McNemars already assumes there is a connection.

Linear Regression

Cleaning

```
pete$Height_cm = as.numeric(gsub("\\$", "", pete$Height_cm))
```

```
## Warning: NAs introduced by coercion
```

```
pete$Armspan_cm = as.numeric(gsub("\\$", "", pete$Armspan_cm))
```

```
## Warning: NAs introduced by coercion
```

```
library(mice)
```

```
## Warning: package 'mice' was built under R version 3.6.3
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
jme <- cbind(Height_cm, Armspan_cm)
set.seed(1909)
imputedat <- mice(jme, m=10, method = "cart")
```

```
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   1   6
##   1   7
##   1   8
##   1   9
##   1   10
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   2   6
##   2   7
##   2   8
##   2   9
##   2   10
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   3   6
##   3   7
##   3   8
##   3   9
##   3   10
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   4   6
##   4   7
##   4   8
##   4   9
##   4   10
##   5   1
##   5   2
##   5   3
##   5   4
##   5   5
##   5   6
##   5   7
##   5   8
##   5   9
##   5   10
```

```
imputedmods <-with(imputedat, lm(Height_cm~Armspan_cm))

summary(pool(imputedmods))
```

```
##          term   estimate  std.error statistic       df p.value
## 1 (Intercept) 26.3188458 2.57752937  10.21088 319.9863       0
## 2  Armspan_cm  0.5103647 0.04175821  12.22190 319.9863       0
```

Slope is 0.5103647 with a standard error of 0.04175821 Intercept is 26.3188458 with a standard error of 2.57752937

Random Forest

```
imputedat2 <- mice(jme, m=10, method = "rf")
```

```
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   1   6
##   1   7
##   1   8
##   1   9
##   1   10
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   2   6
##   2   7
##   2   8
##   2   9
##   2   10
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   3   6
##   3   7
##   3   8
##   3   9
##   3   10
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   4   6
##   4   7
##   4   8
##   4   9
##   4   10
##   5   1
##   5   2
##   5   3
##   5   4
##   5   5
##   5   6
##   5   7
##   5   8
##   5   9
##   5   10
```

```
imputedmods2 <-with(imputedat2, lm(Height_cm~Armspan_cm))

summary(pool(imputedmods2))
```

```
##          term   estimate  std.error statistic       df p.value
## 1 (Intercept) 26.3188458 2.57752937  10.21088 319.9863       0
## 2  Armspan_cm  0.5103647 0.04175821  12.22190 319.9863       0
```

The random forest gave me the same intercept and slope as the cart model

GitHub Link

https://github.com/laeldridge/NP-Homework (https://github.com/laeldridge/NP-Homework)