NC/FET Estrodiol Statistic in R

Matthew Sahagun

5/25/2020

Introduction

For this project, we want to know if estradiol values affect the prgnancy rates of women. To help with this study, below I will clean the data, run some single variable analysis, and then run logistic analysis.

Examining the Data

```
estrodiol = read.csv("/Users/mjs13/Downloads/NC-FET Estradiol ASRM 5-2020 BB (6).csv", header = TRUE)
```

```
head(estrodiol)
```

```
##   Entry.. Data.Access.Group      MRN        Last..First.name Age.at.transfer
## 1      25                   29478195        Khandelwal, Neha              31
## 2     306                   47438858       WEINER, ELIZABETH              37
## 3     305                   34409185         WINGERT, ANGELA              40
## 4      61                   34306183              Wong, Wing              42
## 5      35                   28951499 Poutre, Janeen Michelle              45
## 6     151                   41330333      Chunduri, Poojitha              35
##   Date.of.Transfer Date.of.birth   BMI Race..choice.White.
## 1          5/31/16      12/29/84 27.00           Unchecked
## 2          1/26/18       4/29/80 25.18             Checked
## 3           5/7/18       4/23/78 25.24             Checked
## 4           6/8/17       10/1/74 21.00           Unchecked
## 5         10/22/17       1/24/72 24.00           Unchecked
## 6           5/5/16       4/20/81 24.00           Unchecked
##   Race..choice.South.Asian. Race..choice.East.Asian. Race..choice.Black.or.AA.
## 1                   Checked                Unchecked                 Unchecked
## 2                 Unchecked                Unchecked                 Unchecked
## 3                 Unchecked                Unchecked                 Unchecked
## 4                 Unchecked                  Checked                 Unchecked
## 5                 Unchecked                Unchecked                   Checked
## 6                   Checked                Unchecked                 Unchecked
##   Race..choice.Unknown. Race..choice.Other. Race.other           Ethnicity
## 1             Unchecked           Unchecked            Not hispanic/latino
## 2             Unchecked           Unchecked            Not hispanic/latino
## 3             Unchecked           Unchecked            Not hispanic/latino
## 4             Unchecked           Unchecked            Not hispanic/latino
## 5             Unchecked           Unchecked            Not hispanic/latino
## 6             Unchecked           Unchecked            Not hispanic/latino
##   Smoker SART.Diagnosis..choice.DOR.
## 1  Never                   Unchecked
## 2  Never                   Unchecked
```

```
## 3  Never                   Unchecked
## 4 Former                     Checked
## 5  Never                     Checked
## 6  Never                   Unchecked
##   SART.Diagnosis..choice.Ovulatory.Dysfunction. SART.Diagnosis..choice.PCOS.
## 1                                     Unchecked                   Unchecked
## 2                                     Unchecked                   Unchecked
## 3                                     Unchecked                   Unchecked
## 4                                     Unchecked                   Unchecked
## 5                                     Unchecked                   Unchecked
## 6                                     Unchecked                   Unchecked
##   SART.Diagnosis..choice.RPL. SART.Diagnosis..choice.Male.Factor.
## 1                   Unchecked                           Unchecked
## 2                   Unchecked                           Unchecked
## 3                   Unchecked                           Unchecked
## 4                   Unchecked                           Unchecked
## 5                   Unchecked                           Unchecked
## 6                   Unchecked                           Unchecked
##   SART.Diagnosis..choice.Unexplained. SART.Diagnosis..choice.Endometriosis.
## 1                             Checked                             Unchecked
## 2                             Checked                             Unchecked
## 3                           Unchecked                             Unchecked
## 4                           Unchecked                             Unchecked
## 5                           Unchecked                             Unchecked
## 6                           Unchecked                             Unchecked
##   SART.Diagnosis..choice.Uterine. SART.Diagnosis..choice.Tubal.
## 1                       Unchecked                     Unchecked
## 2                       Unchecked                     Unchecked
## 3                       Unchecked                       Checked
## 4                       Unchecked                     Unchecked
## 5                         Checked                     Unchecked
## 6                       Unchecked                       Checked
##   SART.Diagnosis..choice.Single.gene.disorder. SART.Diagnosis..choice.Other.
## 1                                     Unchecked                   Unchecked
## 2                                     Unchecked                   Unchecked
## 3                                     Unchecked                   Unchecked
## 4                                     Unchecked                   Unchecked
## 5                                     Unchecked                   Unchecked
## 6                                     Unchecked                   Unchecked
##   SART.other Infertility.Type Gravidity Parity..TPAL. Nulliparous
## 1                    Primary         0                         1
## 2                    Primary         1            10           1
## 3                  Secondary         4          1041           0
## 4                    Primary         0                         1
## 5                    Primary         4          1031           0
## 6                  Secondary         3          2012           0
##   Uterine.Cavity.Eval.Type Uterine.Cavity.Eval.Date
## 1                      SIS                 10/9/15
## 2            Hysteroscopy                 10/2/17
## 3            Hysteroscopy                 3/16/18
## 4            Hysteroscopy                  9/2/16
## 5            Hysteroscopy                 8/21/17
## 6                      SIS                 2/22/16
##   Number.of.Embryos.Transferred Catheter.Type Distance.from.Fundus..mm.
```

```
## 1                            1 Cook Echo Tip                         15
## 2                            1       Wallace                         15
## 3                            2       Wallace                         16
## 4                            1 Cook Echo Tip                         15
## 5                            1 Cook Echo Tip                         15
## 6                            1 Cook Echo Tip                         15
##   Cycle.day.of.transfer Endometrial.thickness..mm. Antral.follicle.count
## 1                    17                        8.1                    23
## 2                    19                        8.7                    10
## 3                    22                        8.1                    11
## 4                    19                       12.0                     6
## 5                    23                        7.3                     4
## 6                    22                        7.5                    16
##   Viable.IUP. Ultrasound.details                 Pregnancy.outcomes LB
## 1          No       No pregnancy                       No pregnancy  0
## 2          No        biochemical                        biochemical  0
## 3         Yes  2 IUP @ 7+6, +FCA twin C/S, one female and one male  1
## 4          No       No pregnancy                       No pregnancy  0
## 5          No       No pregnancy                       No pregnancy  0
## 6         Yes           FCA 6w4d                         live birth  1
##   Fertilization.type Embryo.Grade Grade..1.good. Embryo.sex..choice.XX.
## 1                IVF          5BA              1                Checked
## 2              Split          4CC              0              Unchecked
## 3                IVF     4AA, 4BA              1                Checked
## 4               ICSI          6AA              1              Unchecked
## 5               ICSI          6AA              1                Checked
## 6               ICSI          5BB              0                Checked
##   Embryo.sex..choice.XY. Sex.selection.           Age.at.retrieval
## 1              Unchecked             No                         31
## 2                Checked             No                         37
## 3                Checked             No                         39
## 4                Checked             No                         42
## 5              Unchecked             No Patient 44, but used donor
## 6              Unchecked            Yes                         35
##   Number.of.prior.transfers Number.of.prior.failed.transfers Labs.Date LH.value
## 1                         2                                2   5/19/16     6.72
## 2                         1                                1   1/20/18     34.9
## 3                         0                                    4/30/18      7.8
## 4                         0                                     6/3/17     29.6
## 5                         1                                0  10/16/17      162
## 6                         0                                    4/28/18      8.4
##   LH.surge. Cycle.day.of.trigger Estradiol.value Progesterone.value
## 1        No                                61.0               0.48
## 2       Yes                   15           415.3                0.5
## 3        No                   15           184.0                0.3
## 4       Yes                   13           316.0              0.198
## 5       Yes             10/16/17           173.0                1.2
## 6        No                   15           447.0                0.7
##   Anti.mullerian.hormone Vitamin.D.level                  Missing.data
## 1                      3              24 Embyro grade/sex/sex selection
## 2                   4.21              25
## 3                   1.29              38
## 4                    1.7             N/A                          None
## 5                Unknown         Unknown               AMH and Vitamin D
```

```
## 6                   2.11   Not available                              None
##    Complete.
## 1 Incomplete
## 2   Complete
## 3   Complete
## 4   Complete
## 5   Complete
## 6   Complete
```

```r
names(estrodiol)
```

```
##  [1] "Entry.."
##  [2] "Data.Access.Group"
##  [3] "MRN"
##  [4] "Last..First.name"
##  [5] "Age.at.transfer"
##  [6] "Date.of.Transfer"
##  [7] "Date.of.birth"
##  [8] "BMI"
##  [9] "Race..choice.White."
## [10] "Race..choice.South.Asian."
## [11] "Race..choice.East.Asian."
## [12] "Race..choice.Black.or.AA."
## [13] "Race..choice.Unknown."
## [14] "Race..choice.Other."
## [15] "Race.other"
## [16] "Ethnicity"
## [17] "Smoker"
## [18] "SART.Diagnosis..choice.DOR."
## [19] "SART.Diagnosis..choice.Ovulatory.Dysfunction."
## [20] "SART.Diagnosis..choice.PCOS."
## [21] "SART.Diagnosis..choice.RPL."
## [22] "SART.Diagnosis..choice.Male.Factor."
## [23] "SART.Diagnosis..choice.Unexplained."
## [24] "SART.Diagnosis..choice.Endometriosis."
## [25] "SART.Diagnosis..choice.Uterine."
## [26] "SART.Diagnosis..choice.Tubal."
## [27] "SART.Diagnosis..choice.Single.gene.disorder."
## [28] "SART.Diagnosis..choice.Other."
## [29] "SART.other"
## [30] "Infertility.Type"
## [31] "Gravidity"
## [32] "Parity..TPAL."
## [33] "Nulliparous"
## [34] "Uterine.Cavity.Eval.Type"
## [35] "Uterine.Cavity.Eval.Date"
## [36] "Number.of.Embryos.Transferred"
## [37] "Catheter.Type"
## [38] "Distance.from.Fundus..mm."
## [39] "Cycle.day.of.transfer"
## [40] "Endometrial.thickness..mm."
## [41] "Antral.follicle.count"
## [42] "Viable.IUP."
## [43] "Ultrasound.details"
## [44] "Pregnancy.outcomes"
```

```
## [45] "LB"
## [46] "Fertilization.type"
## [47] "Embryo.Grade"
## [48] "Grade..1.good."
## [49] "Embryo.sex..choice.XX."
## [50] "Embryo.sex..choice.XY."
## [51] "Sex.selection."
## [52] "Age.at.retrieval"
## [53] "Number.of.prior.transfers"
## [54] "Number.of.prior.failed.transfers"
## [55] "Labs.Date"
## [56] "LH.value"
## [57] "LH.surge."
## [58] "Cycle.day.of.trigger"
## [59] "Estradiol.value"
## [60] "Progesterone.value"
## [61] "Anti.mullerian.hormone"
## [62] "Vitamin.D.level"
## [63] "Missing.data"
## [64] "Complete."
```

Cleaning the Data

I found this nifty outlier function that seems like it will be useful.

```
#outlier function
outlierKD <- function(dt, var) {
     var_name <- eval(substitute(var),eval(dt))
     na1 <- sum(is.na(var_name))
     m1 <- mean(var_name, na.rm = T)
     par(mfrow=c(2, 2), oma=c(0,0,3,0))
     boxplot(var_name, main="With outliers")
     hist(var_name, main="With outliers", xlab=NA, ylab=NA)
     outlier <- boxplot.stats(var_name)$out
     mo <- mean(outlier)
     var_name <- ifelse(var_name %in% outlier, NA, var_name)
     boxplot(var_name, main="Without outliers")
     hist(var_name, main="Without outliers", xlab=NA, ylab=NA)
     title("Outlier Check", outer=TRUE)
     na2 <- sum(is.na(var_name))
     cat("Outliers identified:", na2 - na1, "n")
     cat("Propotion (%) of outliers:", round((na2 - na1) / sum(!is.na(var_name))*100, 1), "n")
     cat("Mean of the outliers:", round(mo, 2), "n")
     m2 <- mean(var_name, na.rm = T)
     cat("Mean without removing outliers:", round(m1, 2), "n")
     cat("Mean if we remove outliers:", round(m2, 2), "n")
     response <- readline(prompt="Do you want to remove outliers and to replace with NA? [yes/no]: ")
     if(response == "y" | response == "yes"){
          dt[as.character(substitute(var))] <- invisible(var_name)
          assign(as.character(as.list(match.call())$dt), dt, envir = .GlobalEnv)
          cat("Outliers successfully removed", "n")
          return(invisible(dt))
     } else{
          cat("Nothing changed", "n")
          return(invisible(var_name))
```

```
    }
}
```

I determined that all of the values in the column Estradiol.value are factors. I need them to be numeric.

```
#change columns from factor to numeric
estrodiol$Estradiol.value = as.numeric(as.character(estrodiol$Estradiol.value))
class(estrodiol$Estradiol.value)
```

```
## [1] "numeric"
```

```
estrodiol$BMI = as.numeric(as.character(estrodiol$BMI))
class(estrodiol$BMI)
```

```
## [1] "numeric"
```

```
estrodiol$Age.at.transfer = as.numeric(as.character(estrodiol$Age.at.transfer))
class(estrodiol$Age.at.transfer)
```

```
## [1] "numeric"
```

I am checking for missing value. In an earlier version of the data set, there were quite a few instances of missing values. That is no longer the case. I also wanted to look at the boxplot of the data, and noted many outliers.

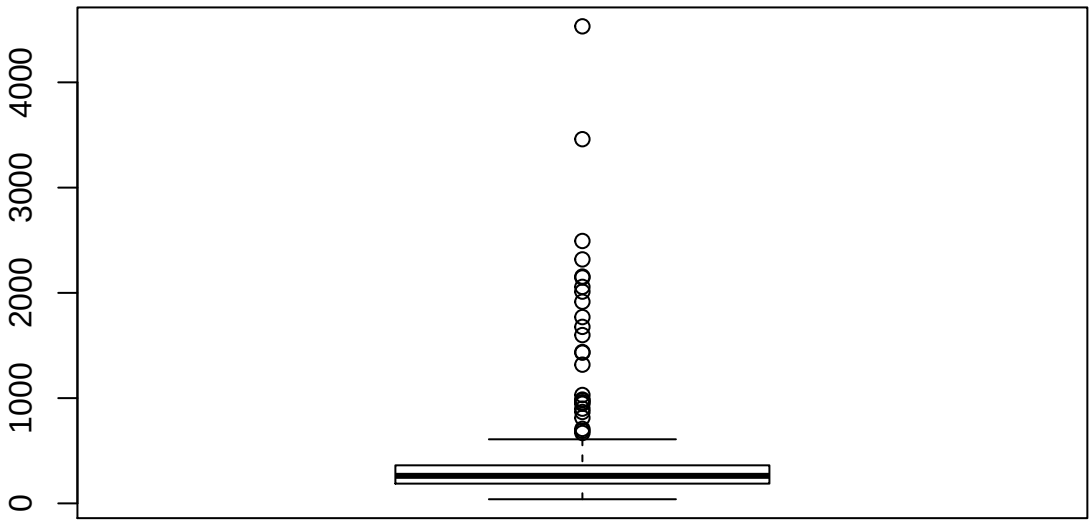
```
#counting the number of NA values
sum(is.na(estrodiol$Estradiol.value))
```

```
## [1] 0
```

```
#removing NA values since they are not relevant to our study.
row = which(is.na(estrodiol$Estradiol.value))
row
```

```
## integer(0)
```

```
#estrodiol = estrodiol[-row,]
#sum(is.na(estrodiol$Estradiol.value))

#boxplot to look at data
boxplot(estrodiol$Estradiol.value)
```




```
#Impute NA from Estrodiol.value column with median
#estrodiol$Estradiol.value[is.na(estrodiol$Estradiol.value)] = median(estrodiol$Estradiol.value, na.rm =

#finding mean of column
mean(estrodiol$Estradiol.value)
```

```
## [1] 346.2098
```

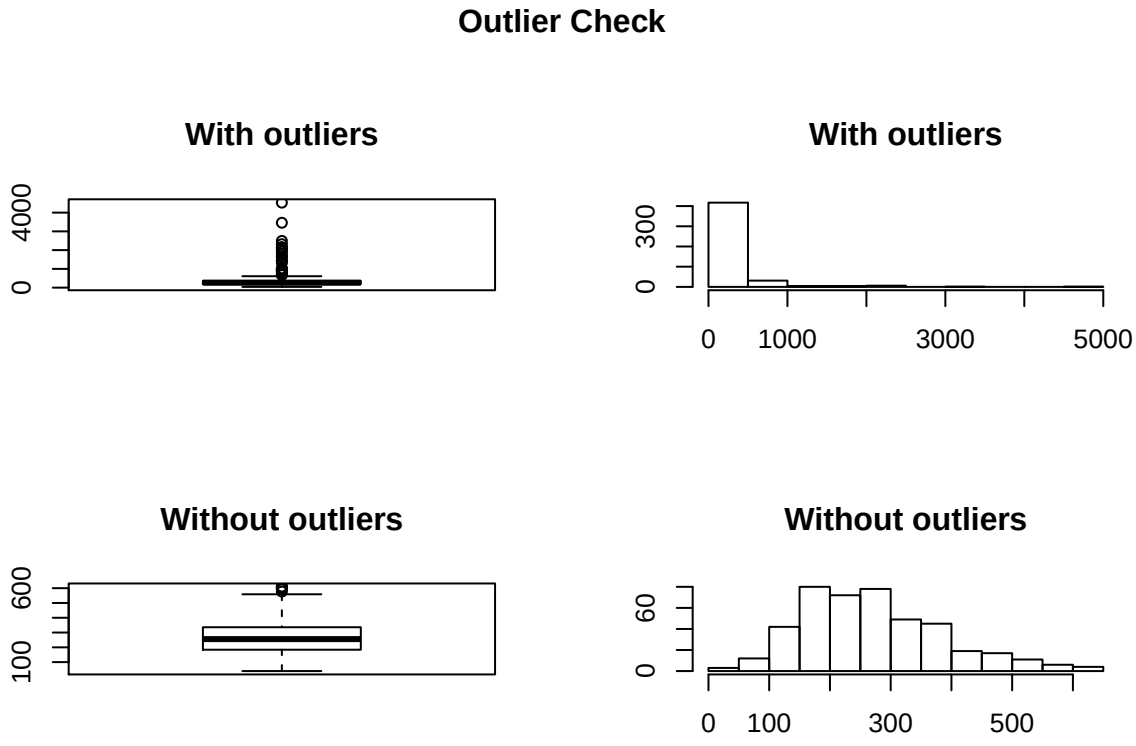
```
#outliers. We have 26 of them
outlier_values = boxplot.stats(estrodiol$Estradiol.value)$out
length(outlier_values)
```

```
## [1] 26
```

```
sort(outlier_values)
```

```
##  [1]  669.0  690.0  708.0  813.0  868.4  900.0  946.5  961.0  979.5  986.0
## [11] 1030.0 1318.0 1432.0 1438.0 1599.0 1677.0 1770.0 1916.0 2012.0 2057.0
## [21] 2146.0 2155.0 2318.0 2493.0 3460.0 4532.0
```

```
#outlier_rows =

outlierKD(estrodiol, Estradiol.value)
```




```
## Outliers identified: 26 nPropotion (%) of outliers: 5.9 nMean of the outliers: 1610.55 nMean without
## Nothing changed n
```

I went in search for the rows of the outliers.

```
#finding the rows of the outliers
row = array()
values = array()

for (i in 1:length(outlier_values)) {
  values[i] = outlier_values[i]
  row[i] = which(estrodiol$Estradiol.value == outlier_values[i])
}

dat = data.frame(row, values)
sorted_dat = dat[order(dat$values),]
sorted_dat
```

```
##    row values
## 21 397  669.0
## 7  162  690.0
## 4   64  708.0
## 20 395  813.0
## 25 456  868.4
## 23 419  900.0
## 24 430  946.5
## 11 289  961.0
```

```
## 18 345  979.5
## 17 333  986.0
## 15 327 1030.0
## 12 312 1318.0
## 19 348 1432.0
## 2   38 1438.0
## 14 323 1599.0
## 5   79 1677.0
## 8  181 1770.0
## 6  108 1916.0
## 26 459 2012.0
## 22 418 2057.0
## 3   61 2146.0
## 16 332 2155.0
## 10 283 2318.0
## 1   24 2493.0
## 13 314 3460.0
## 9  246 4532.0
```

I was told by the researchers that those with an estradiol value greater than 708 were given supplements, which is not part of the study. I vill remove those values, which takes care of my outlier problem.

```
#removing rows where the estrodiol value is greater than 708, since those patients were given supplemen
supp = which(estrodiol$Estradiol.value > 708)
supp
```

```
##  [1]  24  38  61  79 108 181 246 283 289 312 314 323 327 332 333 345 348 395 418
## [20] 419 430 456 459
```

```
length(supp)
```

```
## [1] 23
```

```
estrodiol = estrodiol[-supp,]
nrow(estrodiol)
```

```
## [1] 441
```

```
#Our data now has 441 rows after removing the outliers (greater than 708 estrodiol level)
```

I will also categorize the ages in the following manner, and then add a column for these values into the dataframe: <30, 30-34, 35-39, >=40

```
age_transfer_cat = array()

for (i in 1:nrow(estrodiol)) {
  if (estrodiol$Age.at.transfer[i] < 30) {
    age_transfer_cat[i] = "<30"
  } else if (estrodiol$Age.at.transfer[i] >= 30 & estrodiol$Age.at.transfer[i] < 35) {
    age_transfer_cat[i] = "30-34"
  } else if (estrodiol$Age.at.transfer[i] >= 35 & estrodiol$Age.at.transfer[i] < 40) {
    age_transfer_cat[i] = "35-39"
  } else if (estrodiol$Age.at.transfer[i] >= 40) {
    age_transfer_cat[i] = ">=40"
  }
}

head(age_transfer_cat)
```

```
## [1] "30-34" "35-39" ">=40"  ">=40"  ">=40"  "35-39"
```

```
length(age_transfer_cat)
```

```
## [1] 441
```

```
estrodiol["age_transfer_cat"] = age_transfer_cat
```

Basic Statistics

I will now run some basic statistics.

```
#Now let's run some basic stats
mean((estrodiol$Age.at.transfer))
```

```
## [1] 36.47551
```

```
sd((estrodiol$Age.at.transfer))
```

```
## [1] 3.786512
```

```
mean((estrodiol$Estradiol.value))
```

```
## [1] 273.9999
```

```
sd((estrodiol$Estradiol.value))
```

```
## [1] 120.7039
```

Now we will do some counting statistics. We are going to split our data into groups A and B. Group A has an estradiol level of less than 200.

```
#Now we will do some counting statistics. We are going to split our data into groups A and B. Group A h

row_a = which(estrodiol$Estradiol.value < 200)
group_a = estrodiol[row_a,]
nrow(group_a)
```

```
## [1] 134
```

```
#There are 134 patients in group A

row_b = which(estrodiol$Estradiol.value >= 200)
group_b = estrodiol[row_b,]
nrow(group_b)
```

```
## [1] 307
```

```
#There are 307 patients in group B
```

We will now examine the column called viable IUP. If that columns says yes, that means the patient is pregnant. We want to know the clinical prgnancy rate in total, and for groups A and B.

```
#We will now examine the column called viable IUP. If that columns says yes, that means the patient is

class(estrodiol$Viable.IUP.)
```

```
## [1] "factor"
```

```
unique(estrodiol$Viable.IUP.)
```

```
## [1] No  Yes
## Levels: No Yes
```

```r
yes_rows = which(estrodiol$Viable.IUP. == "Yes")
length(yes_rows)
```

```
## [1] 288
```

```r
#65% of the patients were pregnant

no_rows = which(estrodiol$Viable.IUP. == "No")
length(no_rows)
```

```
## [1] 153
```

```r
#35% of the patients were pregnant

#Now, we want to know what percent of patients from groups A and B had a positive preganacy.
yes_rows_a = which(group_a$Viable.IUP. == "Yes")
length(yes_rows_a)
```

```
## [1] 78
```

```r
#58% of the patients in group A were pregnant

no_rows_a = which(group_a$Viable.IUP. == "No")
length(no_rows_a)
```

```
## [1] 56
```

```r
#42% of the patients in group A were pregnant

yes_rows_b = which(group_b$Viable.IUP. == "Yes")
length(yes_rows_b)
```

```
## [1] 210
```

```r
#68% of the patients in group B were pregnant

no_rows_b = which(group_b$Viable.IUP. == "No")
length(no_rows_b)
```

```
## [1] 97
```

```r
#32% of the patients in group B were pregnant
```

We will know row a chisq test for independence for groups A and B. I will first add groups A and B to the original dataframe

```r
group = array()
for (i in 1:nrow(estrodiol))
  if (estrodiol$Estradiol.value[i] < 200) {
    group[i] = "A"
  } else {
    group[i] = "B"
  }

length(group)
```

```
## [1] 441
```

```
estrodiol["group"] = group
```

To check my work, I will create a contingency table

```
library(MASS)       # load the MASS package
tbl = table(estrodiol$group, estrodiol$Viable.IUP.)
tbl                 # the contingency table
```

```
##
##      No Yes
##   A  56  78
##   B  97 210
```

Now I will run the chisq test

```
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl
## X-squared = 3.8411, df = 1, p-value = 0.05001
```

I will now perform the same operations with live birth columns (LB)

```
#1 corresponds to a live birth. I will first see the total number of 1's
live_birth_rows = (which(estrodiol$LB == 1))
length(live_birth_rows)
```

```
## [1] 264
```

```
#In total, there were 264 live births.

still_birth_rows = (which(estrodiol$LB == 0))
length(still_birth_rows)
```

```
## [1] 170
```

```
#In total, there were 170 non-live births.This does not add up to the total number of rows because some

#Now I will create a contingency table for life births for groups A and B.
tbl2 = table(estrodiol$group, estrodiol$LB)
tbl2                 # the contingency table
```

```
##
##       0   1
##   A  64  69
##   B 106 195
```

```
chisq.test(tbl2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl2
## X-squared = 5.9163, df = 1, p-value = 0.015
```

The findings for live birth for groups A and B are significantly different with an alpha of 0.05

Demographic Statistics

Now we will dive in to some of the demographic statistics.

```
mean_age_a = mean(group_a$Age.at.transfer)
print(paste0("mean age for group A is: ", mean_age_a))
```

```
## [1] "mean age for group A is: 36.0425373134328"
```

```
mean_age_b = mean(group_b$Age.at.transfer)
print(paste0("mean age for group B is: ", mean_age_b))
```

```
## [1] "mean age for group B is: 36.6644951140065"
```

```
mean(estrodiol$BMI)
```

```
## [1] 24.94633
```

```
sd(estrodiol$BMI)
```

```
## [1] 4.776421
```

```
mean_bmi_a = mean(group_a$BMI)
print(paste0("mean BMI for group A is: ", mean_bmi_a))
```

```
## [1] "mean BMI for group A is: 26.1716417910448"
```

```
mean_bmi_b = mean(group_b$BMI)
print(paste0("mean BMI for group B is: ", mean_bmi_b))
```

```
## [1] "mean BMI for group B is: 24.4114983713355"
```

```
mean_end_tot = mean(estrodiol$Endometrial.thickness..mm.)
print(paste0("mean endometrial thickness in total is: ", mean_end_tot))
```

```
## [1] "mean endometrial thickness in total is: 9.11578231292517"
```

```
mean_end_a = mean(group_a$Endometrial.thickness..mm.)
print(paste0("mean endometrial thickness for group A is: ", mean_end_a))
```

```
## [1] "mean endometrial thickness for group A is: 9.20820895522388"
```

```
mean_end_b = mean(group_b$Endometrial.thickness..mm.)
print(paste0("mean endometrial thickness for group B is: ", mean_end_b))
```

```
## [1] "mean endometrial thickness for group B is: 9.07543973941368"
```

```
unique(estrodiol$Smoker)
```

```
## [1] Never  Former
## Levels:  Former Never
```

```
never_smoke_rows = which(estrodiol$Smoker == "Never")
tot_never_smoker = length(never_smoke_rows)
tot_never_smoker
```

```
## [1] 414
```

```
former_smoke_rows = which(estrodiol$Smoker == "Former")
tot_former_smoker = length(former_smoke_rows)
tot_former_smoker
```

```
## [1] 23
```

```r
never_smoke_rows_a = which(group_a$Smoker == "Never")
num_never_smoker_a = length(never_smoke_rows_a)
num_never_smoker_a
```

```
## [1] 125
```

```r
former_smoke_rows_a = which(group_a$Smoker == "Former")
num_former_smoker_a = length(former_smoke_rows_a)
num_former_smoker_a
```

```
## [1] 8
```

```r
never_smoke_rows_b = which(group_b$Smoker == "Never")
num_never_smoker_b = length(never_smoke_rows_b)
num_never_smoker_b
```

```
## [1] 289
```

```r
former_smoke_rows_b = which(group_b$Smoker == "Former")
num_former_smoker_b = length(former_smoke_rows_b)
num_former_smoker_b
```

```
## [1] 15
```

```r
unique(estrodiol$Nulliparous)
```

```
## [1] 1 0
```

```r
null1_rows = which(estrodiol$Nulliparous == 1)
tot_null1 = length(null1_rows)
tot_null1
```

```
## [1] 277
```

```r
null0_rows = which(estrodiol$Nulliparous == 0)
tot_null0 = length(null0_rows)
tot_null0
```

```
## [1] 164
```

```r
null1_rows_a = which(group_a$Nulliparous == 1)
tot_null1_a = length(null1_rows_a)
tot_null1_a
```

```
## [1] 80
```

```r
null0_rows_a = which(group_a$Nulliparous == 0)
tot_null0_a = length(null0_rows_a)
tot_null0_a
```

```
## [1] 54
```

```r
null1_rows_b = which(group_b$Nulliparous == 1)
tot_null1_b = length(null1_rows_b)
tot_null1_b
```

```
## [1] 197
```

```r
null0_rows_b = which(group_b$Nulliparous == 0)
tot_null0_b = length(null0_rows_b)
tot_null0_b
```

```
## [1] 110
```

```
never_smoke_rows_b = which(group_b$Smoker == "Never")
num_never_smoker_b = length(never_smoke_rows_b)
num_never_smoker_b
```

```
## [1] 289
```

```
former_smoke_rows_b = which(group_b$Smoker == "Former")
num_former_smoker_b = length(former_smoke_rows_b)
num_former_smoker_b
```

```
## [1] 15
```

```
unique(estrodiol$Race..choice.White.)
```

```
## [1] Unchecked Checked
## Levels: Checked Unchecked
```

```
unique(estrodiol$Race..choice.East.Asian.)
```

```
## [1] Unchecked Checked
## Levels: Checked Unchecked
```

```
unique(estrodiol$Race..choice.South.Asian.)
```

```
## [1] Checked   Unchecked
## Levels: Checked Unchecked
```

```
tot_white_row = which(estrodiol$Race..choice.White. == "Checked")
length(tot_white_row)
```

```
## [1] 174
```

```
white_rowa = which(group_a$Race..choice.White. == "Checked")
length(white_rowa)
```

```
## [1] 56
```

```
white_rowb = which(group_b$Race..choice.White. == "Checked")
length(white_rowb)
```

```
## [1] 118
```

```
tot_nonwhite_row = which(estrodiol$Race..choice.White. == "Unchecked")
length(tot_nonwhite_row)
```

```
## [1] 267
```

```
nonwhite_rowa = which(group_a$Race..choice.White. == "Unchecked")
length(nonwhite_rowa)
```

```
## [1] 78
```

```
nonwhite_rowb = which(group_b$Race..choice.White. == "Unchecked")
length(nonwhite_rowb)
```

```
## [1] 189
```

```
tot_e_asian_row = which(estrodiol$Race..choice.East.Asian. == "Checked")
length(tot_e_asian_row)
```

```
## [1] 118
```

```
e_asian_rowa = which(group_a$Race..choice.East.Asian. == "Checked")
length(e_asian_rowa)
```

```
## [1] 30
```

```
e_asian_rowb = which(group_b$Race..choice.East.Asian. == "Checked")
length(e_asian_rowb)
```

```
## [1] 88
```

```
tot_none_asian_row = which(estrodiol$Race..choice.East.Asian. == "Unchecked")
length(tot_none_asian_row)
```

```
## [1] 323
```

```
none_asian_rowa = which(group_a$Race..choice.East.Asian. == "Unchecked")
length(none_asian_rowa)
```

```
## [1] 104
```

```
none_asian_rowb = which(group_b$Race..choice.East.Asian. == "Unchecked")
length(none_asian_rowb)
```

```
## [1] 219
```

```
tot_s_asian_row = which(estrodiol$Race..choice.South.Asian. == "Checked")
length(tot_s_asian_row)
```

```
## [1] 106
```

```
s_asian_rowa = which(group_a$Race..choice.South.Asian. == "Checked")
length(s_asian_rowa)
```

```
## [1] 34
```

```
s_asian_rowb = which(group_b$Race..choice.South.Asian. == "Checked")
length(s_asian_rowb)
```

```
## [1] 72
```

```
tot_nons_asian_row = which(estrodiol$Race..choice.South.Asian. == "Unchecked")
length(tot_nons_asian_row)
```

```
## [1] 335
```

```
nons_asian_rowa = which(group_a$Race..choice.South.Asian. == "Unchecked")
length(nons_asian_rowa)
```

```
## [1] 100
```

```
nons_asian_rowb = which(group_b$Race..choice.South.Asian. == "Unchecked")
length(nons_asian_rowb)
```

```
## [1] 235
```

I am also going to add a column into the data set which will be used in the later analysis. Whites and all Asians will be categorized by 1. Non-whites/asians will be categorized by 0.

```
asian_white = array()

for (i in 1:nrow(estrodiol)) {
  if (estrodiol$Race..choice.East.Asian.[i] == "Checked") {
```

```
    asian_white[i] = 1
  } else if (estrodiol$Race..choice.South.Asian.[i] == "Checked") {
    asian_white[i] = 1
  } else if (estrodiol$Race..choice.White.[i] == "Checked") {
    asian_white[i] = 1
  } else {
    asian_white[i] = 0
  }
}

summary(asian_white)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.9025  1.0000  1.0000
```

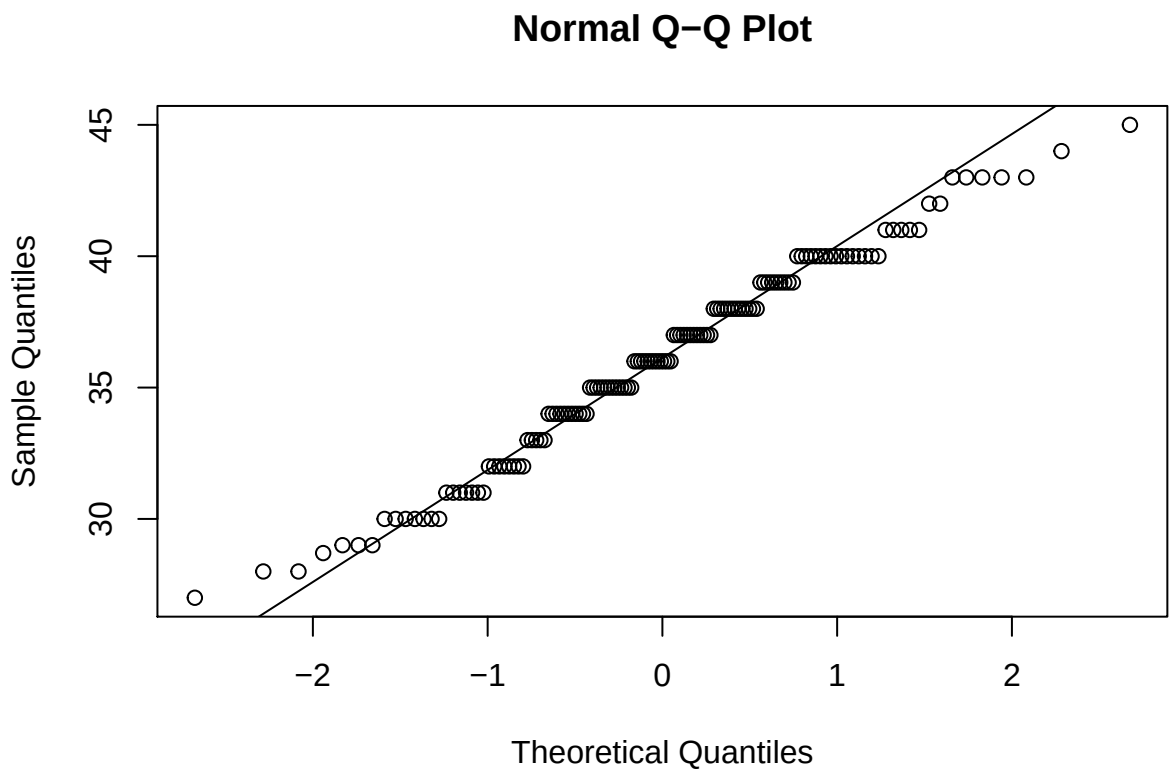
```
sum(asian_white)
```

```
## [1] 398
```

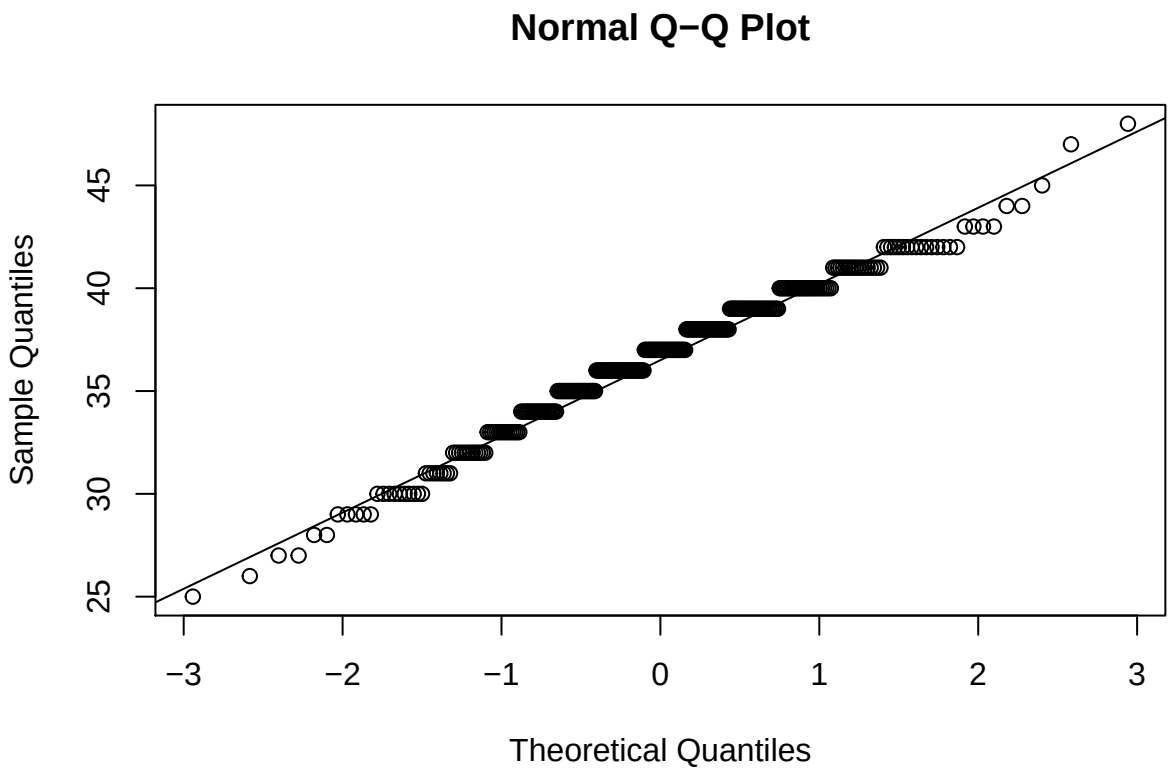
```
estrodiol["asian_white"] = asian_white
```

Now I will run t tests to compare these values for group A and B. I will begin with age.

```
#I will first test to see that these groups are normally distributed.
qqnorm(group_a$Age.at.transfer)
qqline(group_a$Age.at.transfer)
```



```r
qqnorm(group_b$Age.at.transfer)
qqline(group_b$Age.at.transfer)
```

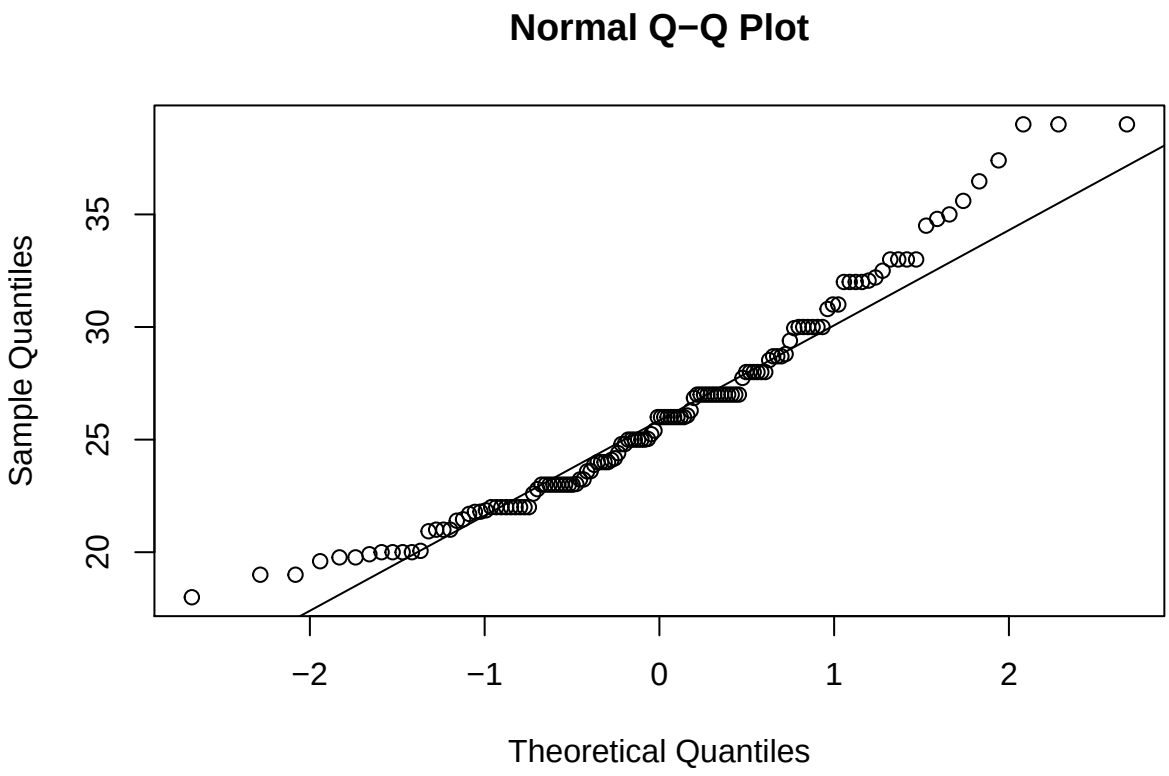



```r
#Both A and B look normal.

t.test(group_a$Age.at.transfer, group_b$Age.at.transfer, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  group_a$Age.at.transfer and group_b$Age.at.transfer
## t = -1.5541, df = 240.93, p-value = 0.1215
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4102910  0.1663754
## sample estimates:
## mean of x mean of y
##  36.04254  36.66450
```

I will continue with BMI.

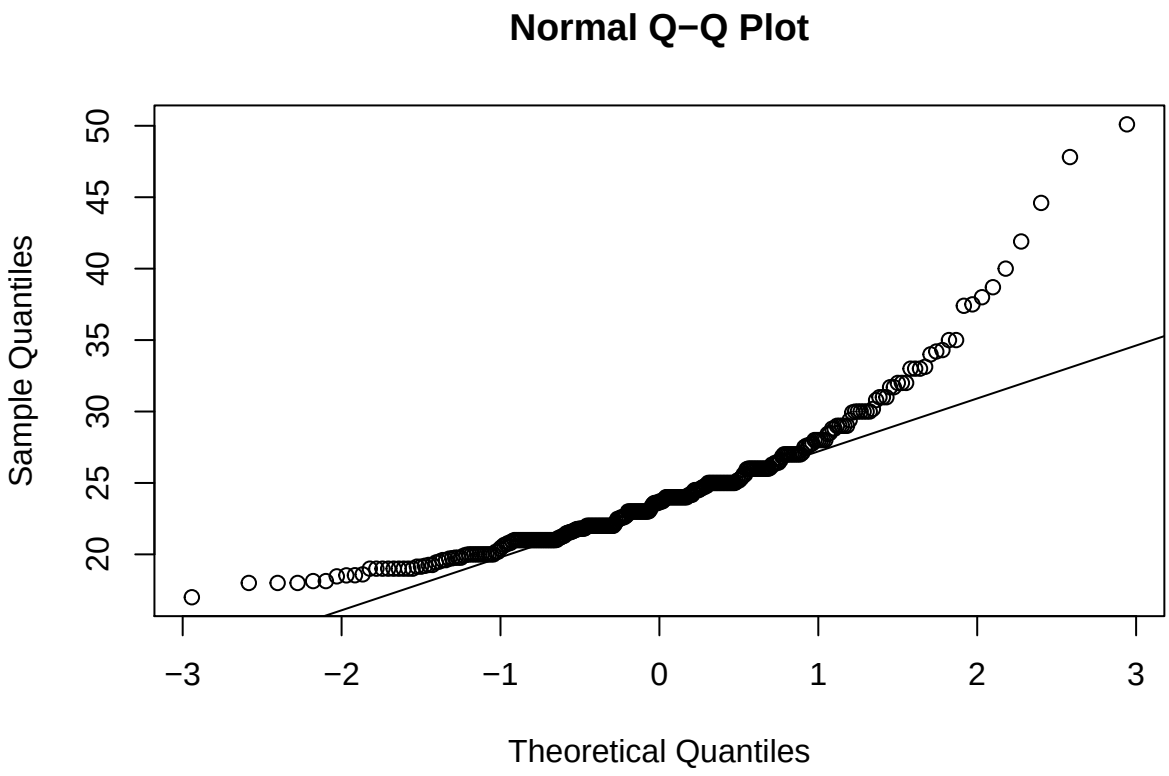
```r
#I will first test to see that these groups are normally distributed.
qqnorm(group_a$BMI)
qqline(group_a$BMI)
```




```
shapiro.test(group_a$BMI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group_a$BMI
## W = 0.95498, p-value = 0.0002199
```

```
qqnorm(group_b$BMI)
qqline(group_b$BMI)
```




```
shapiro.test(group_b$BMI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group_b$BMI
## W = 0.85906, p-value = 4.391e-16
```

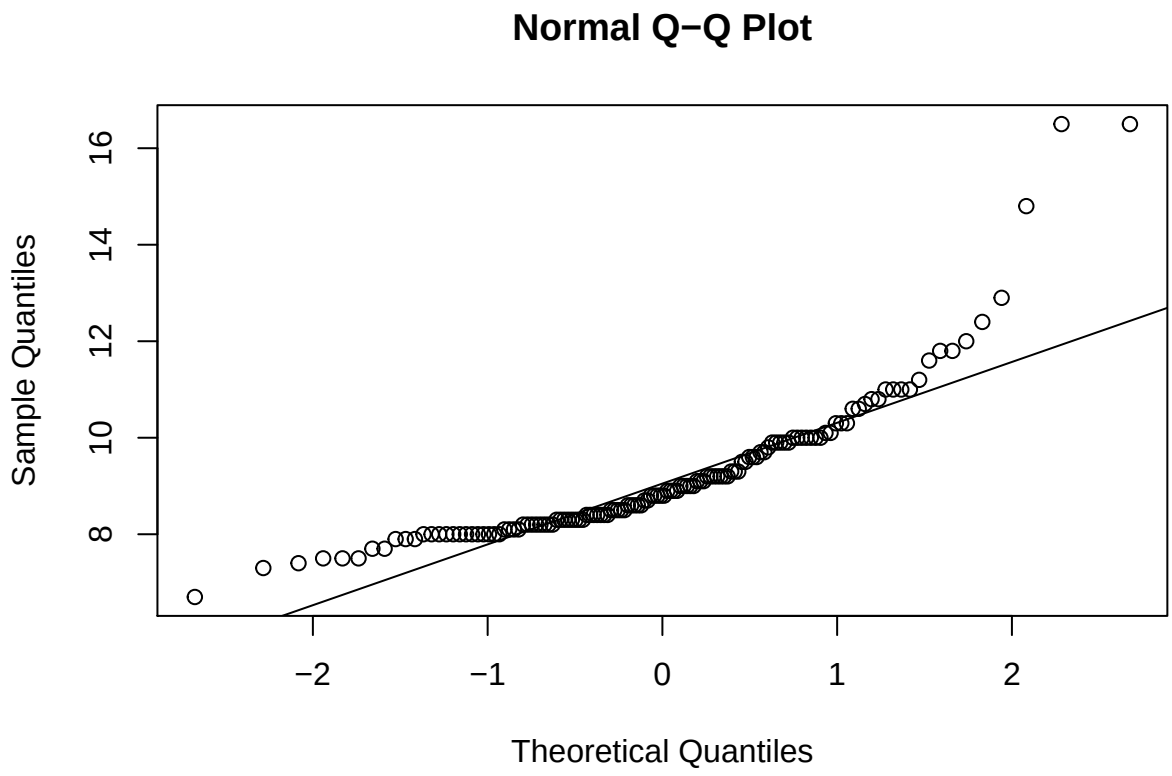
```
#These two groups are definitely not normally distributed. They are very much skewed right.

t.test(group_a$BMI, group_b$BMI, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  group_a$BMI and group_b$BMI
## t = 3.6509, df = 260.64, p-value = 0.0003158
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.810809 2.709478
## sample estimates:
## mean of x mean of y
##  26.17164  24.41150
```

We see from the plots that group B is heavier than group A, which we will control in the logistic regression.

I will continue with endometrial thickness

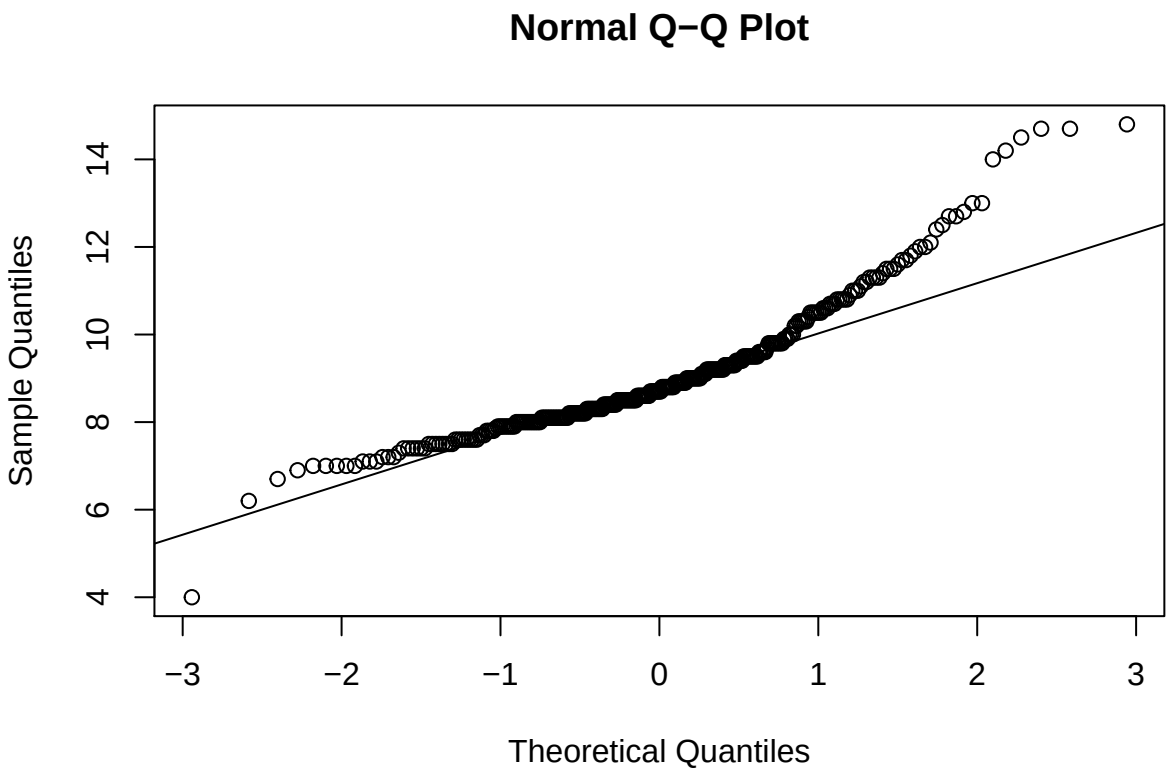
```r
#I will first test to see that these groups are normally distributed.
qqnorm(group_a$Endometrial.thickness..mm.)
qqline(group_a$Endometrial.thickness..mm.)
```




```r
shapiro.test(group_a$Endometrial.thickness..mm.)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group_a$Endometrial.thickness..mm.
## W = 0.8087, p-value = 6.206e-12
```

```r
qqnorm(group_b$Endometrial.thickness..mm.)
qqline(group_b$Endometrial.thickness..mm.)
```




```
shapiro.test(group_b$Endometrial.thickness..mm.)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group_b$Endometrial.thickness..mm.
## W = 0.90952, p-value = 1.269e-12
```

```
#These two groups are definitely not normally distributed. They are very much skewed right.

t.test(group_a$Endometrial.thickness..mm., group_b$Endometrial.thickness..mm., alternative = "two.sided
```

```
##
##  Welch Two Sample t-test
##
## data:  group_a$Endometrial.thickness..mm. and group_b$Endometrial.thickness..mm.
## t = 0.84026, df = 251.99, p-value = 0.4016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1784187  0.4439572
## sample estimates:
## mean of x mean of y
##  9.208209  9.075440
```

The p-value is large so we fail to reject the null hypothesis.

Logistic Regression

Let's now run some multiple logistic regression.

```
glm1 = glm(Viable.IUP. ~ BMI + Endometrial.thickness..mm. +
            Age.at.transfer + Smoker + Estradiol.value
             , data = estrodiol, family = binomial)
summary(glm1)
```

```
##
## Call:
## glm(formula = Viable.IUP. ~ BMI + Endometrial.thickness..mm. +
##     Age.at.transfer + Smoker + Estradiol.value, family = binomial,
##     data = estrodiol)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9249  -1.3356   0.8142   0.9408   1.4179
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  2.1604571  1.7534894   1.232  0.21792
## BMI                          0.0361726  0.0225047   1.607  0.10798
## Endometrial.thickness..mm. -0.0005514  0.0682249  -0.008  0.99355
## Age.at.transfer            -0.0725552  0.0280181  -2.590  0.00961 **
## SmokerFormer               -0.7594099  1.2514173  -0.607  0.54396
## SmokerNever                -0.0712476  1.1793345  -0.060  0.95183
## Estradiol.value             0.0012689  0.0008859   1.432  0.15205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 569.36  on 440  degrees of freedom
## Residual deviance: 555.92  on 434  degrees of freedom
## AIC: 569.92
##
## Number of Fisher Scoring iterations: 4
```

```
glm1 = glm(Viable.IUP. ~ group + BMI + Age.at.transfer, data = estrodiol, family = binomial)
summary(glm1)
```

```
##
## Call:
## glm(formula = Viable.IUP. ~ group + BMI + Age.at.transfer, family = binomial,
##     data = estrodiol)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9843  -1.3218   0.8014   0.9384   1.3390
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.06175    1.16298   1.773  0.07626 .
## groupB           0.57286    0.22274   2.572  0.01011 *
## BMI              0.04088    0.02278   1.795  0.07264 .
```

```
## Age.at.transfer -0.07742    0.02776  -2.789  0.00529 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 569.36  on 440  degrees of freedom
## Residual deviance: 553.99  on 437  degrees of freedom
## AIC: 561.99
##
## Number of Fisher Scoring iterations: 4
```

```
exp(glm1$coefficients) #odds-ratios
```

```
##     (Intercept)          groupB             BMI Age.at.transfer
##       7.8597080       1.7733395       1.0417319       0.9255005
```

```
exp(confint(glm1)) #confidence intervals around the odds-ratios
```

```
## Waiting for profiling to be done...
```

```
##                      2.5 %    97.5 %
## (Intercept)     0.8097448 78.173001
## groupB          1.1460597  2.747981
## BMI             0.9973810  1.090798
## Age.at.transfer 0.8756874  0.976583
```

We see from above that our statistically significant variables are group and age at trasnfer

```
glm2 = glm(Viable.IUP. ~ Age.at.transfer + group +
             BMI + Smoker + Endometrial.thickness..mm. +
             Grade..1.good. + Nulliparous +
             asian_white, data = estrodiol, family = binomial)
summary(glm2)
```

```
##
## Call:
## glm(formula = Viable.IUP. ~ Age.at.transfer + group + BMI + Smoker +
##     Endometrial.thickness..mm. + Grade..1.good. + Nulliparous +
##     asian_white, family = binomial, data = estrodiol)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9357  -1.2876   0.7499   0.9374   1.4935
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   1.60142    1.84732   0.867   0.3860
## Age.at.transfer              -0.06008    0.02947  -2.039   0.0415 *
## groupB                        0.53335    0.22612   2.359   0.0183 *
## BMI                           0.03729    0.02323   1.605   0.1084
## SmokerFormer                 -0.67693    1.27058  -0.533   0.5942
## SmokerNever                  -0.06772    1.19601  -0.057   0.9548
## Endometrial.thickness..mm. -0.01584    0.06923  -0.229   0.8190
## Grade..1.good.                0.37119    0.21345   1.739   0.0820 .
## Nulliparous                   0.29116    0.21959   1.326   0.1849
## asian_white                  -0.23777    0.35846  -0.663   0.5071
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 567.24  on 439  degrees of freedom
## Residual deviance: 544.31  on 430  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 564.31
##
## Number of Fisher Scoring iterations: 4
```

```
exp(glm2$coefficients)
```

```
##                 (Intercept)            Age.at.transfer
##                   4.9600531                  0.9416932
##                      groupB                        BMI
##                   1.7046382                  1.0379973
##                SmokerFormer                SmokerNever
##                   0.5081739                  0.9345266
## Endometrial.thickness..mm.             Grade..1.good.
##                   0.9842836                  1.4494541
##                 Nulliparous                asian_white
##                   1.3379820                  0.7883851
```

```
exp(confint(glm2))
```

```
## Waiting for profiling to be done...
```

```
##                                   2.5 %     97.5 %
## (Intercept)                 0.14845202 258.262919
## Age.at.transfer             0.88819365   0.997215
## groupB                      1.09402447   2.658495
## BMI                         0.99295947   1.087905
## SmokerFormer                0.02206579   5.132825
## SmokerNever                 0.04418391   8.003290
## Endometrial.thickness..mm. 0.86045249   1.130111
## Grade..1.good.              0.95285801   2.202227
## Nulliparous                 0.86850428   2.056372
## asian_white                 0.37890803   1.560682
```

After some back and forth, the researcher wished to study the response viable IUP with the following predictors. The only change from above is that we broke the age ranges into groups, so age is no longer a continuous variable.

```
glm3 = glm(Viable.IUP. ~ age_transfer_cat + group +
             BMI + Smoker + Endometrial.thickness..mm. +
             Grade..1.good. + Nulliparous +
             asian_white, data = estrodiol, family = binomial)
summary(glm3)
```

```
##
## Call:
## glm(formula = Viable.IUP. ~ age_transfer_cat + group + BMI +
##     Smoker + Endometrial.thickness..mm. + Grade..1.good. + Nulliparous +
##     asian_white, family = binomial, data = estrodiol)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0153  -1.2744   0.7347   0.9398   1.5382
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.40123    1.59875  -0.251   0.8018
## age_transfer_cat>=40        -0.20990    0.56759  -0.370   0.7115
## age_transfer_cat30-34        0.56148    0.57345   0.979   0.3275
## age_transfer_cat35-39       -0.21191    0.54425  -0.389   0.6970
## groupB                       0.55251    0.22765   2.427   0.0152 *
## BMI                          0.03690    0.02345   1.574   0.1156
## SmokerFormer                -0.86846    1.28769  -0.674   0.5000
## SmokerNever                 -0.27092    1.21328  -0.223   0.8233
## Endometrial.thickness..mm. -0.01252    0.07056  -0.177   0.8591
## Grade..1.good.               0.39329    0.21437   1.835   0.0666 .
## Nulliparous                  0.25769    0.22023   1.170   0.2420
## asian_white                 -0.22432    0.35945  -0.624   0.5326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 567.24  on 439  degrees of freedom
## Residual deviance: 540.14  on 428  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 564.14
##
## Number of Fisher Scoring iterations: 4
```

```
exp(glm3$coefficients)
```

```
##                (Intercept)       age_transfer_cat>=40
##                  0.6694958                  0.8106625
##      age_transfer_cat30-34      age_transfer_cat35-39
##                  1.7532692                  0.8090346
##                     groupB                        BMI
##                  1.7376171                  1.0375934
##               SmokerFormer                SmokerNever
##                  0.4195961                  0.7626808
## Endometrial.thickness..mm.             Grade..1.good.
##                  0.9875559                  1.4818424
##                Nulliparous                asian_white
##                  1.2939335                  0.7990617
```

```
exp(confint(glm3))
```

```
## Waiting for profiling to be done...
```

```
##                                  2.5 %    97.5 %
## (Intercept)                 0.03283954 23.775419
## age_transfer_cat>=40        0.25111099  2.397951
## age_transfer_cat30-34       0.53833182  5.261628
## age_transfer_cat35-39       0.26053907  2.277167
## groupB                      1.11215835  2.718870
## BMI                         0.99216014  1.087985
```

```
## SmokerFormer               0.01788966  4.419559
## SmokerNever                0.03542445  6.815591
## Endometrial.thickness..mm. 0.86092042  1.136498
## Grade..1.good.             0.97264601  2.256111
## Nulliparous                0.83895698  1.991315
## asian_white                0.38354624  1.585918
```

Now we will repeat this analysis, but with live birth as the response.

```
glm4 = glm(LB ~ age_transfer_cat + group +
             BMI + Smoker + Endometrial.thickness..mm. +
             Grade..1.good. + Nulliparous +
             asian_white, data = estrodiol, family = binomial)
summary(glm4)
```

```
##
## Call:
## glm(formula = LB ~ age_transfer_cat + group + BMI + Smoker +
##     Endometrial.thickness..mm. + Grade..1.good. + Nulliparous +
##     asian_white, family = binomial, data = estrodiol)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9338  -1.2225   0.7574   0.9986   1.6615
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.95828    1.57691  -0.608  0.54339
## age_transfer_cat>=40       -0.38621    0.56963  -0.678  0.49777
## age_transfer_cat30-34       0.59299    0.57441   1.032  0.30191
## age_transfer_cat35-39      -0.20869    0.54697  -0.382  0.70281
## groupB                      0.66387    0.22558   2.943  0.00325 **
## BMI                         0.03423    0.02281   1.500  0.13349
## SmokerFormer               -1.27197    1.27921  -0.994  0.32006
## SmokerNever                -0.50003    1.19740  -0.418  0.67624
## Endometrial.thickness..mm.  0.05411    0.07081   0.764  0.44474
## Grade..1.good.              0.36963    0.21262   1.738  0.08214 .
## Nulliparous                 0.02319    0.21936   0.106  0.91581
## asian_white                -0.04129    0.34587  -0.119  0.90498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 579.25  on 432  degrees of freedom
## Residual deviance: 549.09  on 421  degrees of freedom
##   (8 observations deleted due to missingness)
## AIC: 573.09
##
## Number of Fisher Scoring iterations: 4
```

```
exp(glm4$coefficients)
```

```
##                (Intercept)       age_transfer_cat>=40
##                  0.3835515                  0.6796307
##      age_transfer_cat30-34      age_transfer_cat35-39
```

```
##                   1.8093849                   0.8116488
##                      groupB                         BMI
##                   1.9422920                   1.0348250
##                SmokerFormer                SmokerNever
##                   0.2802789                   0.6065145
## Endometrial.thickness..mm.             Grade..1.good.
##                   1.0556050                   1.4471951
##                 Nulliparous                asian_white
##                   1.0234619                   0.9595535
```

```
exp(confint(glm4))
```

```
## Waiting for profiling to be done...
```

```
##                                  2.5 %    97.5 %
## (Intercept)                0.01955188 13.143818
## age_transfer_cat>=40       0.21040209  2.028418
## age_transfer_cat30-34      0.55666798  5.469809
## age_transfer_cat35-39      0.26127798  2.312764
## groupB                     1.25032301  3.031749
## BMI                        0.99052699  1.083504
## SmokerFormer               0.01204602  2.883706
## SmokerNever                0.02866190  5.226239
## Endometrial.thickness..mm. 0.92042889  1.216243
## Grade..1.good.             0.95351450  2.196525
## Nulliparous                0.66404505  1.570745
## asian_white                0.47874737  1.873523
```