# Final-Report

Anh Nguyen

2022-12-18

# Introduction

## Background

Obesity has been listed as one of the global health problems in Western countries and, together with obesity-caused cardiovascular diseases, are among the most dangerous diseases in the world. Genders and ethnicity has reportedly been factors influencing the severity of this obesity, more specifically higher prevalence is found in female than in male, and in African Americans and Hispanic American over other groups.

## Motivation

Controlling overweight and identifying efficient treatments for obesity are of paramount importance.

## Justification to conduct the study

This RCT study was carried out to determine the efficiency of a behavioral weight loss program for obese individuals at different intensities (low and moderate). The study also aimed to investigate different factors that influence the amount of weight loss, including but not limited to gender, age, ethnicity, intervention duration, and tobacco and alcohol consumption.

## Scientific questions of interest

- How much weight loss do two groups (intervention and control) achieve? Do intervention group lose more weight than the control group do?
- Which factors significantly influence the weight loss outcomes and which factors do not?
- Does intervention duration correlate with the weight loss outcomes?

# Data Description

In this study, participants were randomized to 2 groups – an intervention group (IG) which were exposed to a moderate behavioral weight loss program and a control group (CG) which underwent the same program with low intensity. The demographic data were collected once for each participant, including clinical site, height, age, gender, race/ethnicity, number of obesity-related comorbid conditions, education, marital status, current tobacco use, and current alcohol use. Participants' weight were recorded pre- and post-intervention, as well as number of months they participated in the program. There were 261 participants enrolled (124 in

IG and 137 in CG), but there were 95 records missing information of final weight and length of intervention. That means only 166 participants completed the program (76 in IG and 90 in CG). We will use these 166 records for data analysis.

The primary outcome of this study is the amount of weight loss, which is the difference between baseline and final weight of each individual – this is a continuous variable. The secondary outcomes are the time from randomization to final weight – this can be both a continuous variable and an ordinal categorical variable – and the status of weight loss of a participant, either losing or gaining weight – this is a dichotomous categorical variable.

The candidates for predictors of the weight loss outcomes are gender, race/ethnicity, and baseline weight. We are also going to test the predictive values of the other independent variables, such as age, clinical site, and current tobacco and alcohol use.

In the below analysis, we are going to use significance level of 0.05.

### Summary of demographic variables

```
table1(~Site + Gender + age + height + Race + comorbid + Education + Marital + Tobacco +
    Alcohol | Condition, data = included.data, topclass = table1.styles, overall = F,
    extra.col = list(`P-value` = pvalue))
```

```
## Warning in chisq.test(table(y, g)): Chi-squared approximation may be incorrect
```

```
## Get nicer 'table1' LaTeX output by simply installing the 'kableExtra' package
```

| | Control | Intervention | P-value |
|---|---|---|---|
| | (N=90) | (N=76) | |
| Clinical site | | | |
| Site 1 | 23 (25.6%) | 15 (19.7%) | 0.642 |
| Site 2 | 23 (25.6%) | 21 (27.6%) | |
| Site 3 | 19 (21.1%) | 21 (27.6%) | |
| Site 4 | 13 (14.4%) | 7 (9.2%) | |
| Site 5 | 12 (13.3%) | 12 (15.8%) | |
| Gender | | | |
| Male | 14 (15.6%) | 12 (15.8%) | 1 |
| Female | 76 (84.4%) | 64 (84.2%) | |
| Age (years) | | | |
| Mean (SD) | 48.8 (11.1) | 49.1 (11.2) | 0.85 |
| Median [Min, Max] | 51.0 [18.0, 68.0] | 51.0 [21.0, 68.0] | |
| Height (cm) | | | |
| Mean (SD) | 165 (8.75) | 164 (7.97) | 0.937 |
| Median [Min, Max] | 164 [147, 187] | 164 [149, 183] | |
| Race/ethnicity | | | |
| Asian | 0 (0%) | 2 (2.6%) | 0.323 |
| African American (non-Hispanic black) | 59 (65.6%) | 54 (71.1%) | |
| Hispanic/Latino | 13 (14.4%) | 9 (11.8%) | |
| Non-Hispanic white | 18 (20.0%) | 11 (14.5%) | |
| Number of obesity-related comorbid conditions | | | |
| Mean (SD) | 1.09 (1.16) | 1.29 (1.08) | 0.251 |
| Median [Min, Max] | 1.00 [0, 5.00] | 1.00 [0, 4.00] | |

| | Control | Intervention | P-value |
|---|---|---|---|
| Education | | | |
| <= 12 years | 33 (36.7%) | 18 (23.7%) | 0.102 |
| > 12 years | 57 (63.3%) | 58 (76.3%) | |
| Marital status | | | |
| Not married | 47 (52.2%) | 45 (59.2%) | 0.456 |
| Married | 43 (47.8%) | 31 (40.8%) | |
| Current tobacco use | | | |
| No | 74 (82.2%) | 63 (82.9%) | 1 |
| Yes | 16 (17.8%) | 13 (17.1%) | |
| Current alcohol use | | | |
| No | 63 (70.0%) | 53 (69.7%) | 1 |
| Yes | 27 (30.0%) | 23 (30.3%) | |

There is no significant difference in the demographic variables between 2 groups.

**Summary of intervention-related and outcome variables**

```
table1(~wgt_base + wgt_final + month + wgt_change + Losing_Wgt_Status | Condition,
    data = included.data, topclass = table1.styles, overall = F, extra.col = list(`P-value` = pvalue))
```

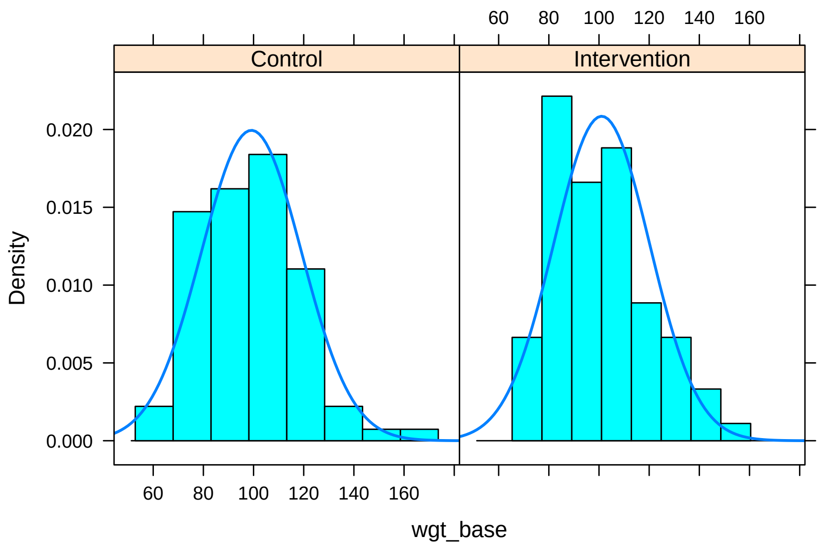
```
## Get nicer 'table1' LaTeX output by simply installing the 'kableExtra' package
```

| | Control | Intervention | P-value |
|---|---|---|---|
| | (N=90) | (N=76) | |
| Baseline weight (kg) | | | |
| Mean (SD) | 99.2 (20.1) | 101 (19.3) | 0.536 |
| Median [Min, Max] | 98.5 [57.4, 163] | 100 [68.2, 151] | |
| Final weight (kg) | | | |
| Mean (SD) | 98.4 (20.7) | 98.6 (19.1) | 0.945 |
| Median [Min, Max] | 97.3 [54.8, 163] | 96.7 [59.4, 151] | |
| Time since randomization to final weight (months) | | | |
| Mean (SD) | 11.8 (1.21) | 11.7 (1.48) | 0.517 |
| Median [Min, Max] | 11.0 [10.0, 15.0] | 11.0 [10.0, 17.0] | |
| Weight change after intervention (kg) | | | |
| Mean (SD) | -0.794 (4.14) | -2.48 (4.93) | 0.0196 |
| Median [Min, Max] | -0.900 [-13.7, 11.5] | -1.80 [-16.0, 8.10] | |
| Losing_Wgt_Status | | | |
| Losing Weight | 55 (61.1%) | 53 (69.7%) | 0.318 |
| Not Losing Weight | 35 (38.9%) | 23 (30.3%) | |

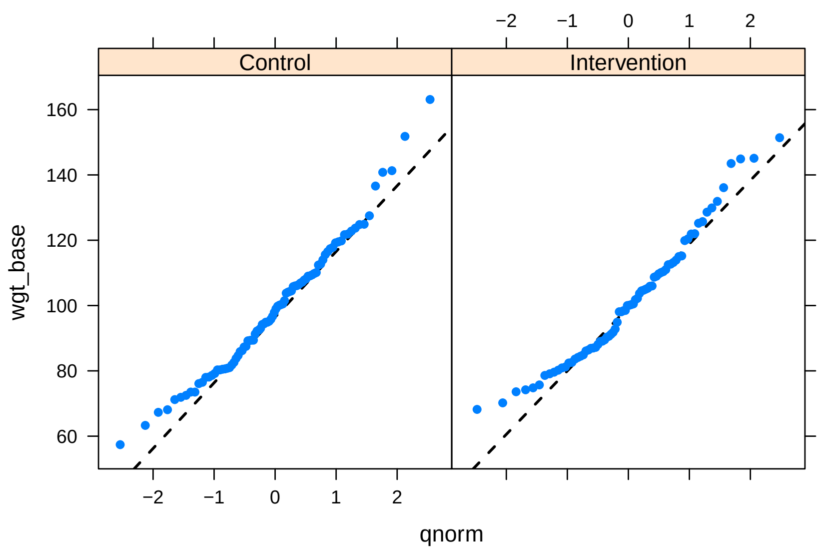
For weight change status, those with negative weight change (final weight minus baseline weight) are considered as "Losing Weight", and others are "Not Losing Weight". We assume this outcome variable follows a binomial distribution. An important assumption needs to be made is the independence of outcome variables among participants.

**Potential distribution of outcome variables**

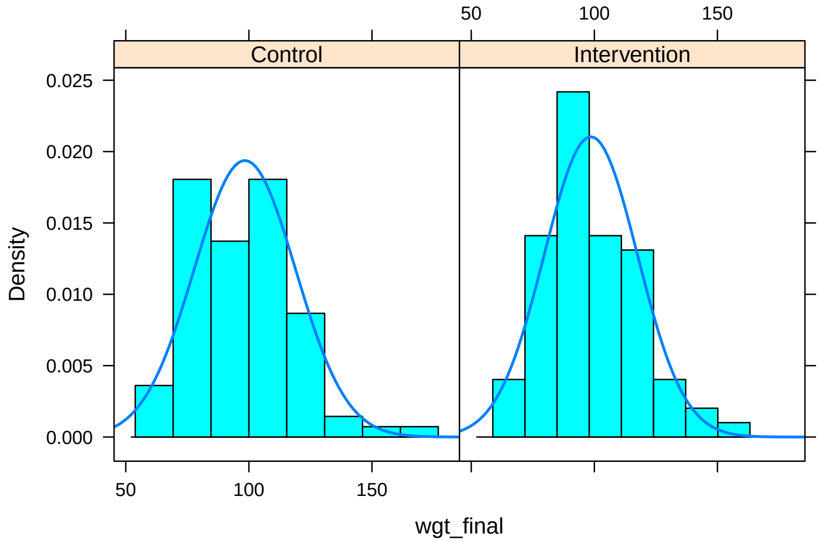
```
histogram(~wgt_base | Condition, data = included.data, fit = "normal")
```



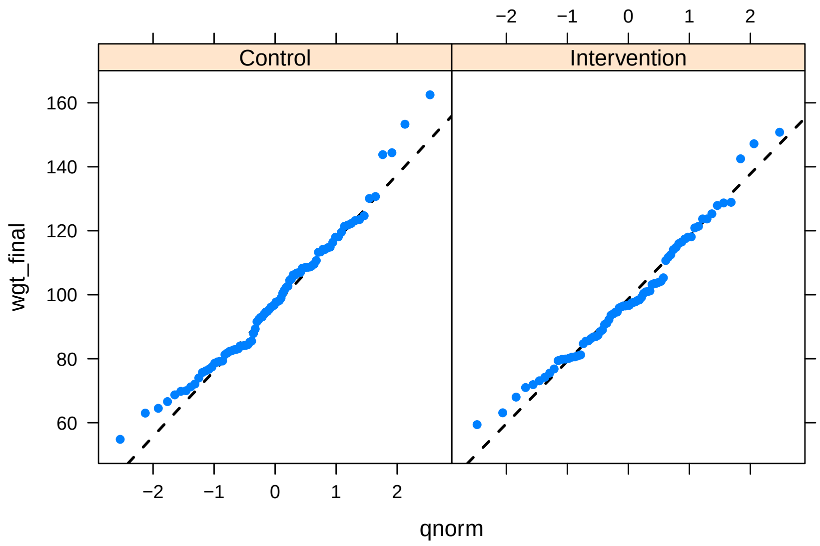

```
xqqmath(~wgt_base | Condition, data = included.data)
```



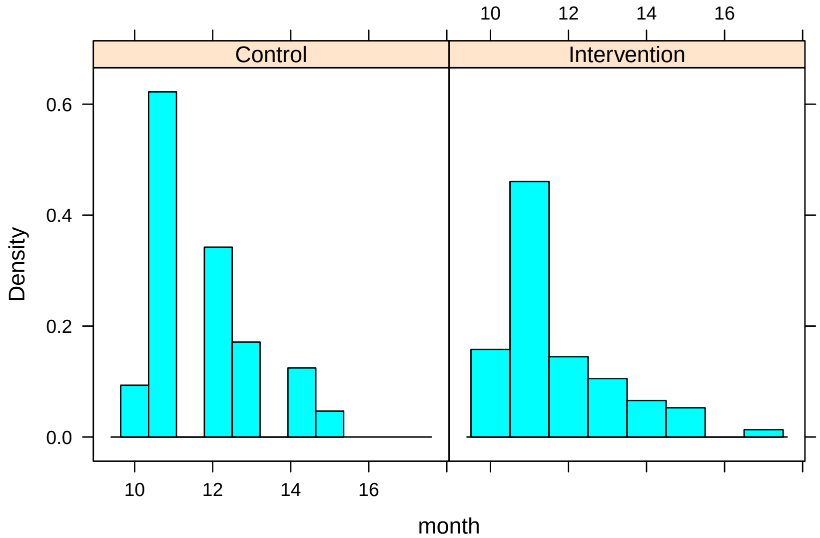

```
histogram(~wgt_final | Condition, data = included.data, fit = "normal")
```



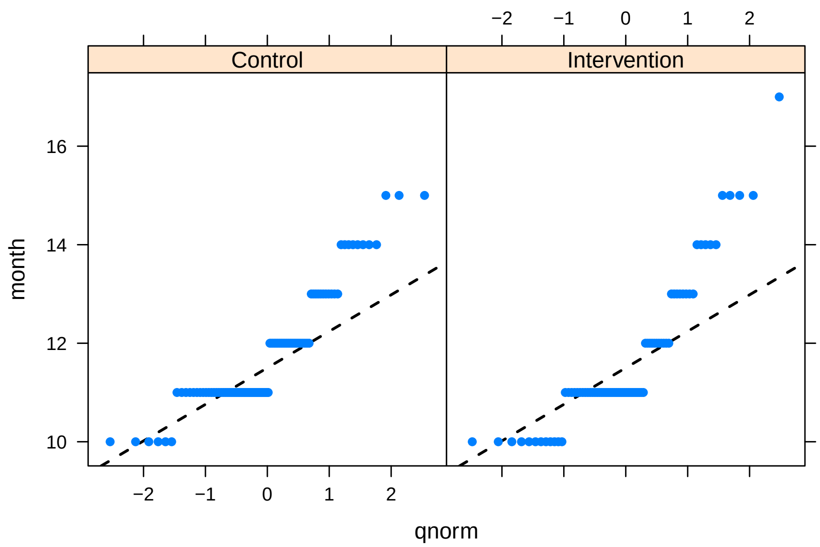

```r
xqqmath(~wgt_final | Condition, data = included.data)
```



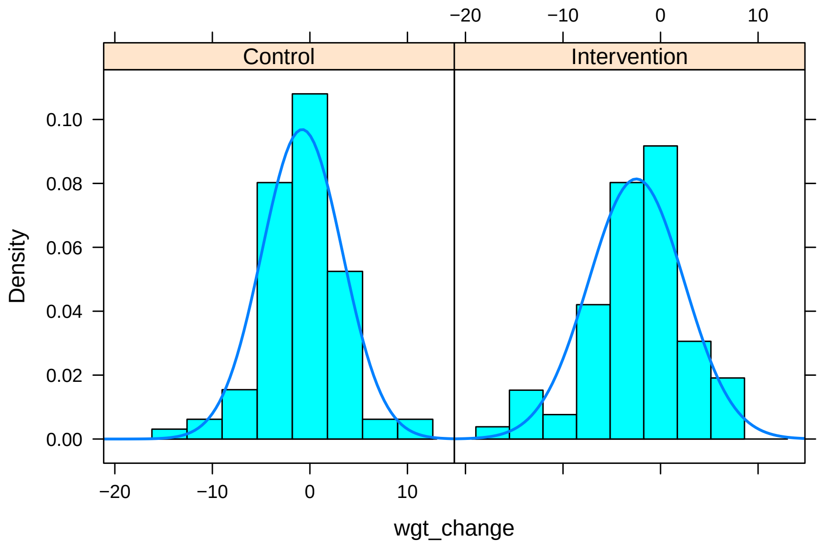

```r
histogram(~month | Condition, data = included.data)
```



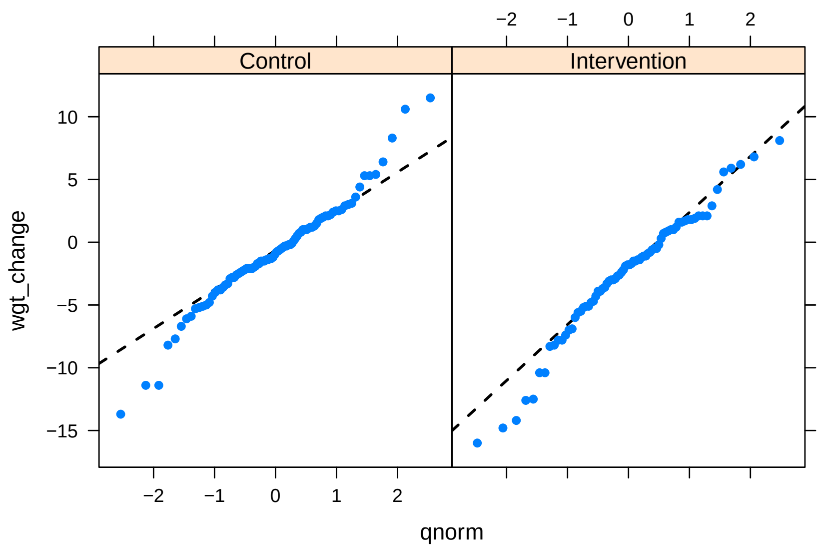

```r
xqqmath(~month | Condition, data = included.data)
```




```r
histogram(~wgt_change | Condition, data = included.data, fit = "normal")
```




```
xqqmath(~wgt_change | Condition, data = included.data)
```




From the Q-Q plots for baseline weight, final weight, and weight change variables, it seems off in the tails, however not by much. I think it is safe to assume the normal distribution for these variables. For number of months of intervention, we will use non-parametric analysis methods.

# Point Estimates and Confidence Intervals

## Obtain point estimates for population parameters of interest.

Sample means of the continuous variables, which are estimation for their population means, can be found in the summary tables in the section above. Here we calculate 95% CI for the means of the continuous variables that are assumed to be normal.

```
list_cont <- list("age", "height", "wgt_base", "wgt_final", "wgt_change")

for (label_cont in list_cont) {
    print(paste("====", label_cont, "===="))
    # print(t.test(x=included.data[[label__cont]], conf.level = 0.95)$conf.int)
    print("Control")
    print(t.test(x = included.data[included.data$tx == "0", ][[label_cont]], conf.level = 0.95)$conf.in
    print("Intervention")
    print(t.test(x = included.data[included.data$tx == "1", ][[label_cont]], conf.level = 0.95)$conf.in
}
```

```
## [1] "==== age ===="
## [1] "Control"
## [1] 46.46229 51.09327
## [1] "Intervention"
## [1] 46.54961 51.66092
## [1] "==== height ===="
## [1] "Control"
## [1] 162.7642 166.4291
## [1] "Intervention"
## [1] 162.6728 166.3141
## [1] "==== wgt_base ===="
## [1] "Control"
## [1]  94.94667 103.37111
## [1] "Intervention"
## [1]  96.65578 105.45738
## [1] "==== wgt_final ===="
## [1] "Control"
## [1]  94.02763 102.70126
## [1] "Intervention"
## [1]  94.21441 102.93822
## [1] "==== wgt_change ===="
## [1] "Control"
## [1] -1.66173334  0.07284445
## [1] "Intervention"
## [1] -3.607630 -1.352897
```

For ordinal and non-parametric variables such as “comorbid” and “month”, we use the bootstrap method to calculate 95% for their means.

```
list.bootstrap = list("month", "comorbid")
for (i in list.bootstrap) {
    m <- sapply(X = 1:10000, FUN = function(X) {
        mean(sample(x = included.data[included.data$tx == "0", ][[i]], size = nrow(included.data[includ
            "0", ]), replace = TRUE))
    })
    # hist(x = m, freq=FALSE, main = paste('Sampling distribution of', i))
    print(paste("====", i, "===="))
    print("Control")
    print(quantile(x = m, probs = c(0.025, 0.975)))
    m <- sapply(X = 1:10000, FUN = function(X) {
        mean(sample(x = included.data[included.data$tx == "1", ][[i]], size = nrow(included.data[includ
            "0", ]), replace = TRUE))
```

```
    })
    # hist(x = m, freq=FALSE, main = paste('Sampling distribution of', i))
    print("Intervention")
    print(quantile(x = m, probs = c(0.025, 0.975)))
}
```

```
## [1] "==== month ===="
## [1] "Control"
##     2.5%    97.5%
## 11.57778 12.06667
## [1] "Intervention"
##     2.5%    97.5%
## 11.38889 12.00000
## [1] "==== comorbid ===="
## [1] "Control"
##      2.5%     97.5%
## 0.8555556 1.3444444
## [1] "Intervention"
##     2.5%    97.5%
## 1.066667 1.522222
```

For variables following binomial distribution (gender, education, martial status, current tobacco and alcohol use, and losing weight status), we calculate the 95% CI for the binom proportion (for "male", "<= 12 years", "not married", "no" for smoking and drinking, and "losing weight", respectively).

```
list_binom <- list("Gender", "Education", "Marital", "Tobacco", "Alcohol", "Losing_Wgt_Status")

for (label in list_binom) {
    print(paste("====", label, "===="))
    # print(t.test(x=included.data[[label__cont]], conf.level = 0.95)$conf.int)
    print("Control")
    print(prop.test(x = included.data[included.data$tx == "0", ][[label]], n = nrow(included.data[inclu
        "0", ]), correct = FALSE)$conf.int[1:2] * 100)
    print("Intervention")
    print(prop.test(x = included.data[included.data$tx == "1", ][[label]], n = nrow(included.data[inclu
        "1", ]), correct = FALSE)$conf.int[1:2] * 100)
}
```

```
## [1] "==== Gender ===="
## [1] "Control"
## [1]  9.498276 24.432845
## [1] "Intervention"
## [1]  9.269497 25.601432
## [1] "==== Education ===="
## [1] "Control"
## [1] 27.44725 46.97770
## [1] "Intervention"
## [1] 15.53933 34.36138
## [1] "==== Marital ===="
## [1] "Control"
## [1] 42.02456 62.23794
## [1] "Intervention"
```

```
## [1] 47.97855 69.55620
## [1] "==== Tobacco ===="
## [1] "Control"
## [1] 73.05611 88.75026
## [1] "Intervention"
## [1] 72.90213 89.72198
## [1] "==== Alcohol ===="
## [1] "Control"
## [1] 59.87349 78.48909
## [1] "Intervention"
## [1] 58.66580 78.90867
## [1] "==== Losing_Wgt_Status ===="
## [1] "Control"
## [1] 50.78246 70.53008
## [1] "Intervention"
## [1] 58.66580 78.90867
```

Most variables show overlapping 95% CI of mean between IG and CG. Variables such as marital status, education, and comorbid shows difference in 95% CIs of mean, however no significant difference in sample means between IG and CG are found.

Weight change and Losing weight status are the only two variables showing big difference in 95% CIs of mean between 2 groups.

# Hypothesis Testing

## Weight changes pre and post intervention

### Are weight losses in 2 group significant?

The alternative hypotheses for 2 groups are mean weights in both groups at pre and post-intervention are different. We use paired t-test for each sample.

```
t.test(x = included.data[included.data$tx == 0, ]$wgt_final, y = included.data[included.data$tx ==
    0, ]$wgt_base, paired = TRUE, mu = 0, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  included.data[included.data$tx == 0, ]$wgt_final and included.data[included.data$tx == 0, ]$wg
## t = -1.8201, df = 89, p-value = 0.07211
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.66173334  0.07284445
## sample estimates:
## mean difference
##      -0.7944444
```

```
t.test(x = included.data[included.data$tx == 1, ]$wgt_final, y = included.data[included.data$tx ==
    1, ]$wgt_base, paired = TRUE, mu = 0, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  included.data[included.data$tx == 1, ]$wgt_final and included.data[included.data$tx == 1, ]$wg
## t = -4.3827, df = 75, p-value = 3.758e-05
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -3.607630 -1.352897
## sample estimates:
## mean difference
##       -2.480263
```

The amount of weight change in CG is not significant (mean = -0.794, p=0.072), while IG shows significant weight loss (mean = -2.48, p = 3.8e-5).

### Do intervention group lose more weight than the control group do?

In the summary table in "Data Description" section, a two-sided t test has been used to compare the difference between weight change between 2 groups. The null hypothesis is 2 groups have the similar weight change after the program. The result from t-test rejects this H0.

We can conduct a one-sided t-test with the same H0, and H1 is that weight loss in IG is greater than in CG. First, we test whether sample variance in weight change between 2 groups are equal:

```
var.test(x = included.data[included.data$tx == 0, ]$wgt_change, y = included.data[included.data$tx ==
    1, ]$wgt_change, ratio = 1, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  included.data[included.data$tx == 0, ]$wgt_change and included.data[included.data$tx == 1, ]$w
## F = 0.70447, num df = 89, denom df = 75, p-value = 0.1128
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4529449 1.0866522
## sample estimates:
## ratio of variances
##          0.7044708
```

Variance in weight change between 2 groups is not significantly different (p=0.113)

Now, assuming equal population variance in weight change between 2 groups, we compare the mean weight change between 2 groups using one.sided test:

```
t.test(x = included.data[included.data$tx == 1, ]$wgt_change, y = included.data[included.data$tx ==
    0, ]$wgt_change, mu = 0, var.equal = TRUE, alternative = "less", conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  included.data[included.data$tx == 1, ]$wgt_change and included.data[included.data$tx == 0, ]$w
## t = -2.3938, df = 164, p-value = 0.008903
```

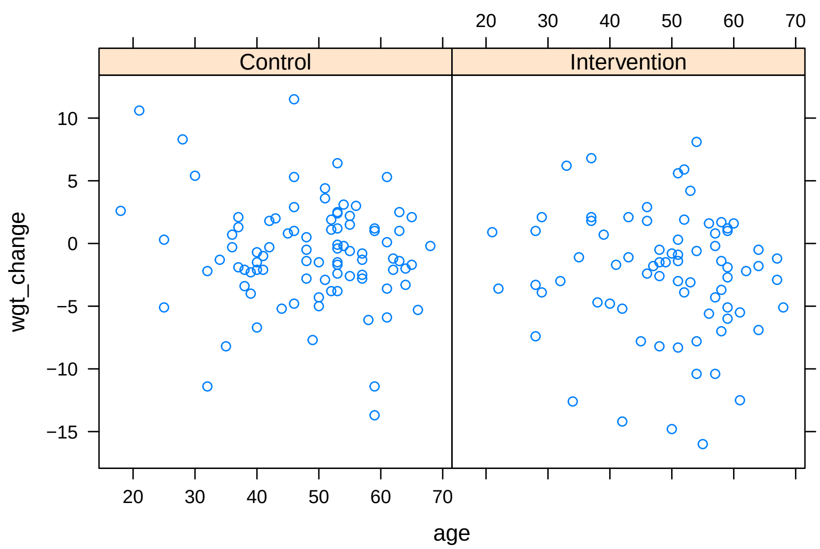
```
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5208466
## sample estimates:
##  mean of x  mean of y
## -2.4802632 -0.7944444
```

Weight loss in intervention group is significantly greater than in CG (p=0.009).

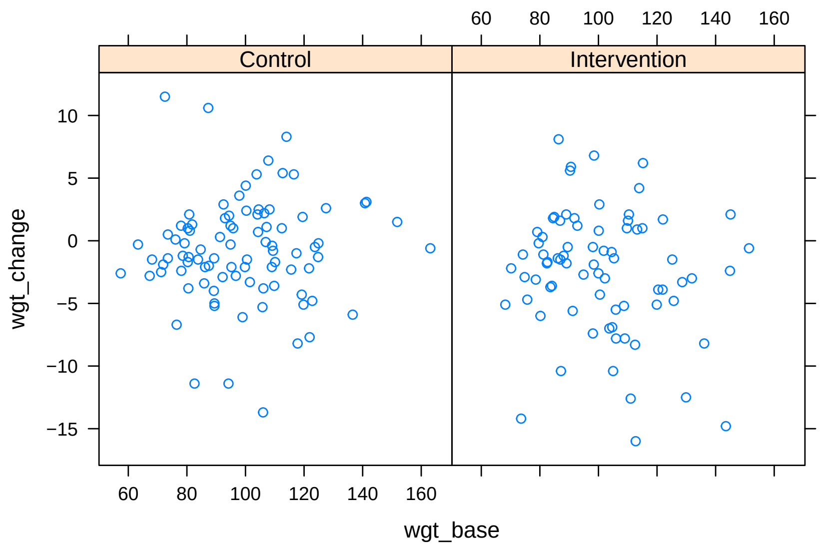
## Linear Model

We can start by plotting the linear relationship between amount of weight change after program and other continuous variables such as age, height, and baseline weight.

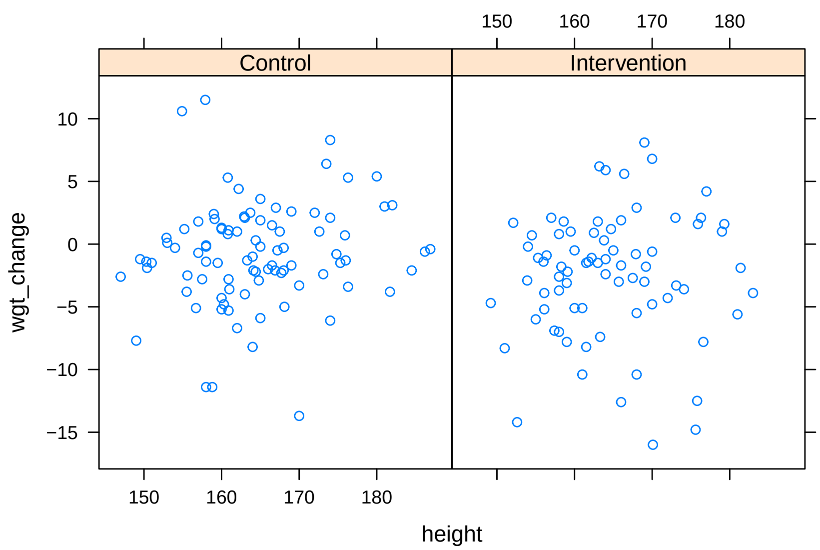
```
xyplot(wgt_change ~ age | Condition, data = included.data)
```




```
xyplot(wgt_change ~ wgt_base | Condition, data = included.data)
```




```
xyplot(wgt_change ~ height | Condition, data = included.data)
```




### Multiple linear regression model

From these plots, we can not observe any linear relationship. Now let's fit a multiple linear regression model for these continuous variables for 2 groups and perform ANOVA tests to test the H0 which is all $\beta$ equal 0 or weight change linearly depends on 3 continuous variables.

```
(linear.m0 <- lm(wgt_change ~ wgt_base + age + height, data = included.data[included.data$tx ==
    0, ], x = TRUE))
```

```
##
```

```
## Call:
## lm(formula = wgt_change ~ wgt_base + age + height, data = included.data[included.data$tx ==
##     0, ], x = TRUE)
##
## Coefficients:
## (Intercept)     wgt_base          age       height
##   -8.621523    -0.008445    -0.047194     0.066626
```

```
anova(linear.m0)
```

```
## Analysis of Variance Table
##
## Response: wgt_change
##           Df  Sum Sq Mean Sq F value Pr(>F)
## wgt_base   1    2.80  2.8001  0.1631 0.6873
## age        1   25.12 25.1155  1.4633 0.2297
## height     1   22.04 22.0369  1.2839 0.2603
## Residuals 86 1476.11 17.1641
```

```
(linear.m1 <- lm(wgt_change ~ wgt_base + age + height, data = included.data[included.data$tx ==
    1, ], x = TRUE))
```

```
##
## Call:
## lm(formula = wgt_change ~ wgt_base + age + height, data = included.data[included.data$tx ==
##     1, ], x = TRUE)
##
## Coefficients:
## (Intercept)     wgt_base          age       height
##    -4.70516     -0.05921     -0.06440      0.06913
```

```
anova(linear.m1)
```

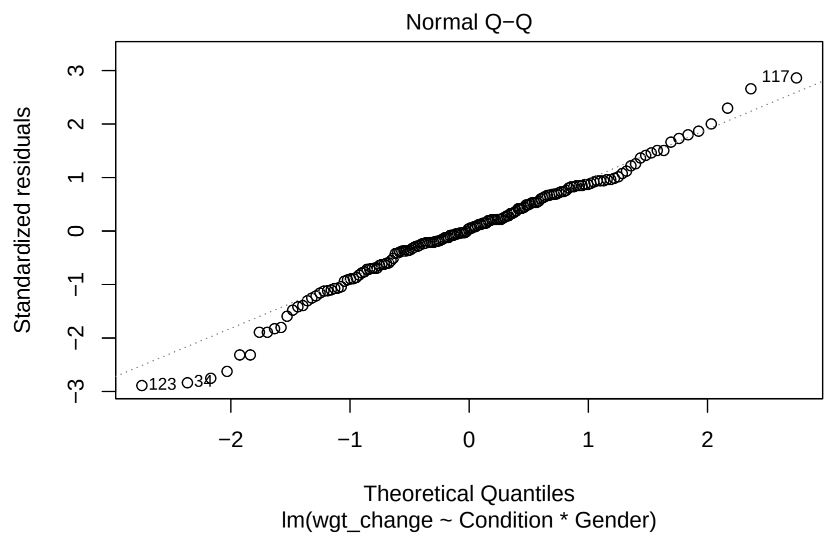
```
## Analysis of Variance Table
##
## Response: wgt_change
##           Df  Sum Sq Mean Sq F value Pr(>F)
## wgt_base   1   48.17  48.167  2.0172 0.1598
## age        1   37.91  37.905  1.5874 0.2118
## height     1   20.20  20.199  0.8459 0.3608
## Residuals 72 1719.23  23.878
```

AVOVA tests fail to reject H0, meaning that weight change does not significantly have linear relationship with age, height and baseline weight in both groups. This can also be interpreted that these variables are not predictors for post-intervention weight loss .

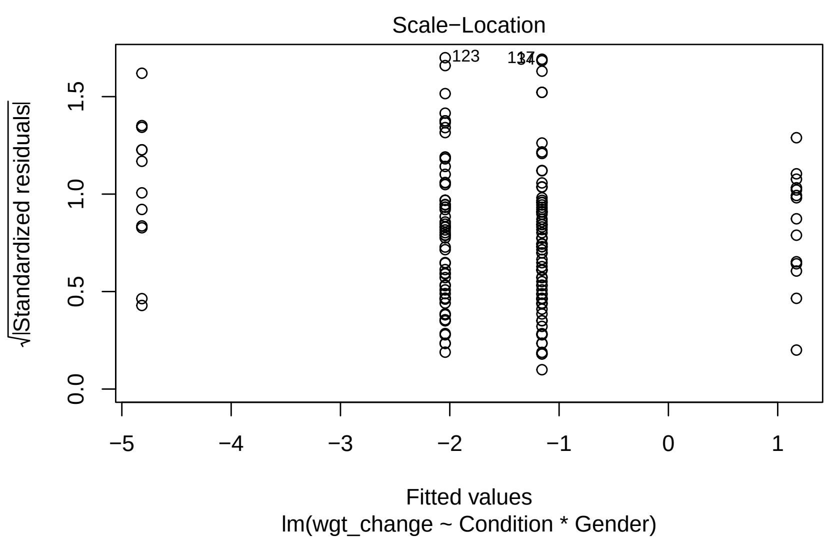
### Categorical independent variables

We use two-way ANOVA tests to predict weight change as a function of intervention condition and one of categorical variables, including clinical site, gender, race/ethnicity, number of comorbid conditions, number of months of intervention, education, marital status, and current tobacco and alcohol use.

```r
mod = lm(wgt_change ~ Condition * Gender, data = included.data)
plot(mod, which = 2)
```




```r
plot(mod, add.smooth = FALSE, which = 3)
```

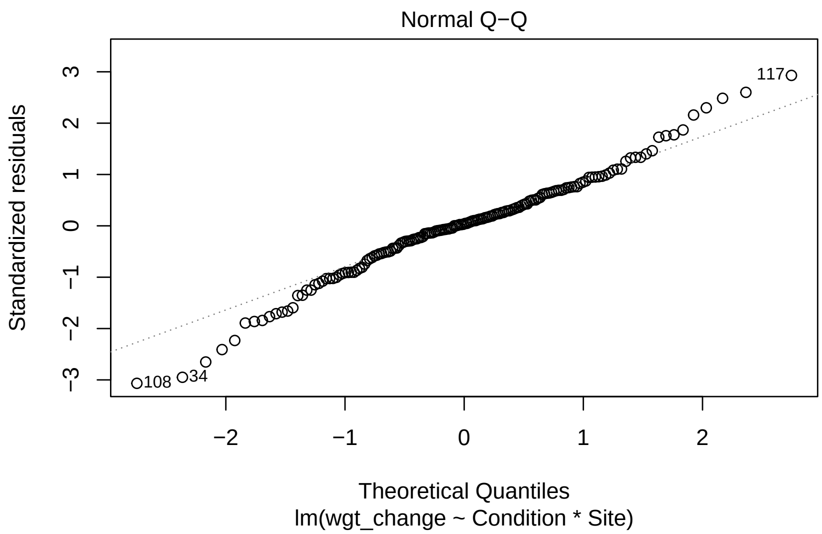



```r
vwgt_chgn_aov <- aov(mod)
summary(vwgt_chgn_aov)
```

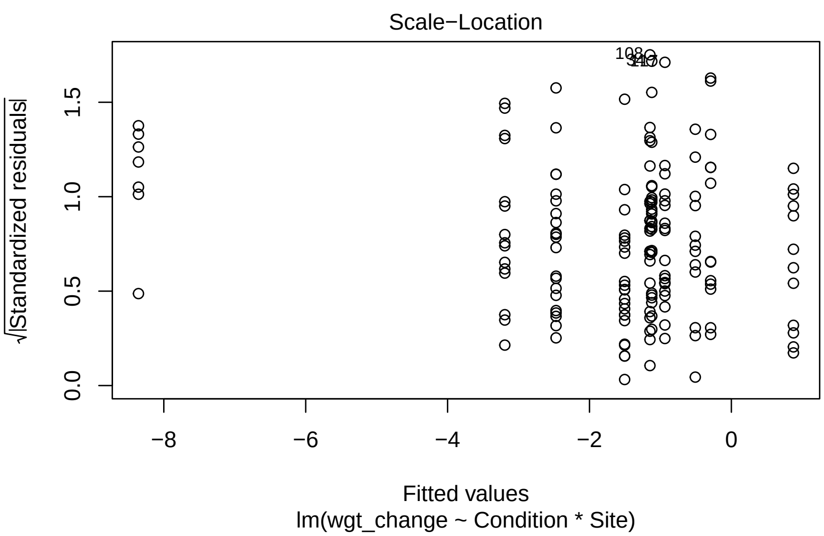
```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## Condition          1    117  117.10   5.910 0.01614 *
## Gender             1      0    0.01   0.001 0.98033
## Condition:Gender   1    142  141.85   7.159 0.00822 **
```

```
## Residuals        162   3210   19.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod = lm(wgt_change ~ Condition * Site, data = included.data)
plot(mod, which = 2)
```




```
plot(mod, add.smooth = FALSE, which = 3)
```

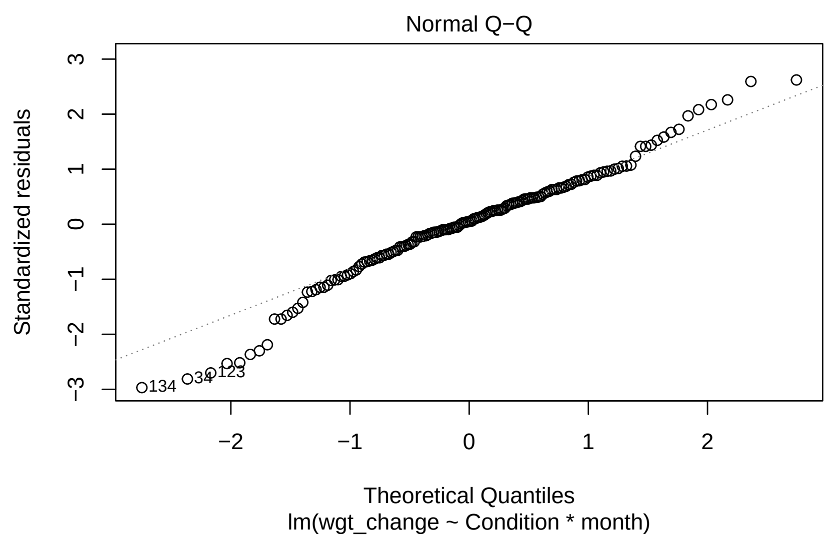



```
vwgt_chgn_aov <- aov(mod)
summary(vwgt_chgn_aov)
```

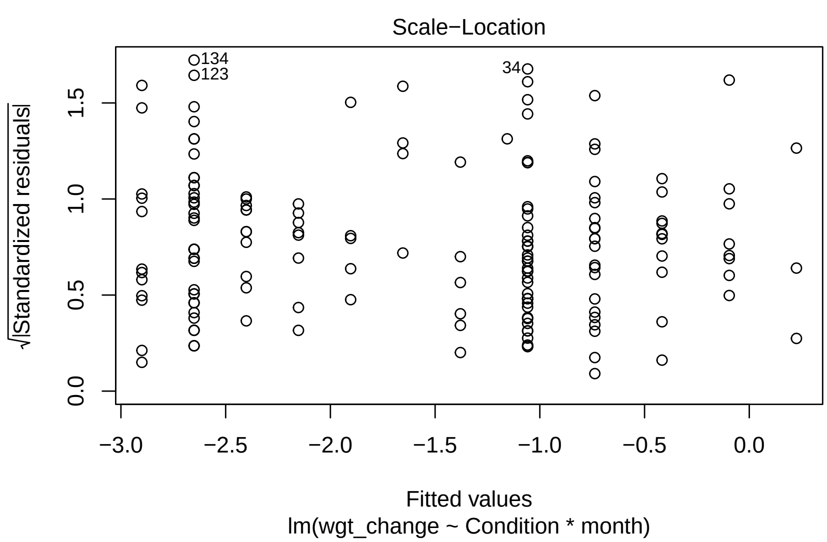
```
##                    Df Sum Sq Mean Sq F value Pr(>F)
```

```
## Condition          1   117.1  117.10    6.157 0.0141 *
## Site               4   171.7   42.93    2.257 0.0654 .
## Condition:Site     4   212.8   53.20    2.797 0.0280 *
## Residuals        156 2967.1   19.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod = lm(wgt_change ~ Condition * month, data = included.data)
plot(mod, which = 2)
```




```
plot(mod, add.smooth = FALSE, which = 3)
```



```r
vwgt_chgn_aov <- aov(mod)
summary(vwgt_chgn_aov)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## Condition         1    117  117.10   5.701 0.0181 *
## month             1     23   23.34   1.136 0.2880
## Condition:month   1      0    0.38   0.018 0.8927
## Residuals       162   3328   20.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use Q-Q plots and residual-versus-fit plots to check the assumptions of normality and homogeneity of variances. The two-way ANOVA tests for each model incorporating one of the categorical variables show the significant effect of interaction between intervention condition and gender (p=0.008) and interaction between intervention condition and clinical site (p=0.028). Other factors, such as months after randomization, do not show any significant effect on weight change.

## Categorical Data Analysis

We will construct multiple contingency tables displaying the relationship between losing weight status and one of the categorical variables. Only the categorical variables found with significant relationship will be shown below.

```r
addmargins(table(included.data$Site, included.data$Losing_Wgt_Status))
```

```
##
##          Losing Weight Not Losing Weight Sum
##   Site 1            24                14  38
##   Site 2            36                 8  44
##   Site 3            25                15  40
##   Site 4            13                 7  20
##   Site 5            10                14  24
##   Sum              108                58 166
```

```r
chisq.test(table(included.data$Site, included.data$Losing_Wgt_Status))$expected
```

```
##
##          Losing Weight Not Losing Weight
##   Site 1      24.72289         13.277108
##   Site 2      28.62651         15.373494
##   Site 3      26.02410         13.975904
##   Site 4      13.01205          6.987952
##   Site 5      15.61446          8.385542
```

```r
# no cell have expected value <5 => use Chi-squared test for association

chisq.test(table(included.data$Site, included.data$Losing_Wgt_Status))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(included.data$Site, included.data$Losing_Wgt_Status)
## X-squared = 11.389, df = 4, p-value = 0.02252
```

```r
(p.site1 <- 24/38)
```

```
## [1] 0.6315789
```

```r
(p.site2 <- 36/44)
```

```
## [1] 0.8181818
```

```r
(p.site3 <- 25/40)
```

```
## [1] 0.625
```

```r
(p.site4 <- 13/20)
```

```
## [1] 0.65
```

```r
(p.site5 <- 10/24)
```

```
## [1] 0.4166667
```

```r
prop.test(x = c(36, 10), n = c(44, 24), p = NULL, alternative = "two.sided", correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of c36 out of 4410 out of 24
## X-squared = 9.6782, df = 1, p-value = 0.001865
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1415215 0.6615088
## sample estimates:
##    prop 1    prop 2
## 0.8181818 0.4166667
```

Chi-squared tests were used for most contingency tables, except for the 4x2 contingency table between losing weight status and race/ethnicity (approximation was not valid, so Fisher's exact test was applied).

We found only the significant relationship between clinical site and losing weight status (p=0.023). Post-hoc two-proportion z-test shows that proportions of people losing weight in different sites are significant different (p=0.002). Most people in site 2 lost weight, while least people in site 5 lost weight.

## Construct a simple generalized linear regression model.

First we construct the full version of generalized linear model for the binary outcome (Losing weight status) for each group

```r
print("==== Control group ====")
```

```
## [1] "==== Control group ===="
```

```r
glm_wgt_chng <- glm(Losing_Wgt ~ Gender + age + Site + height + wgt_base + month +
    comorbid + Education + Marital + Tobacco + Alcohol, family = binomial, data = included.data[included
    "0", ])
summary(glm_wgt_chng)
```

```
##
## Call:
## glm(formula = Losing_Wgt ~ Gender + age + Site + height + wgt_base +
##     month + comorbid + Education + Marital + Tobacco + Alcohol,
##     family = binomial, data = included.data[included.data$tx ==
##         "0", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9059  -0.9520   0.3436   0.8510   1.9645
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.453696   9.946209  -0.247  0.80514
## GenderFemale        2.267674   1.381423   1.642  0.10068
## age                 0.018378   0.025300   0.726  0.46760
## SiteSite 2          1.682403   0.908332   1.852  0.06400 .
## SiteSite 3          0.860613   0.751893   1.145  0.25238
## SiteSite 4         -0.287222   1.062172  -0.270  0.78684
## SiteSite 5         -1.587686   1.085832  -1.462  0.14369
## height             -0.005435   0.059040  -0.092  0.92666
## wgt_base            0.006319   0.018992   0.333  0.73936
## month              -0.047565   0.240243  -0.198  0.84305
## comorbid           -0.150981   0.257931  -0.585  0.55831
## Education> 12 years -0.011081   0.597143  -0.019  0.98519
## MaritalMarried      0.062003   0.560280   0.111  0.91188
## TobaccoYes          3.195182   1.213840   2.632  0.00848 **
## AlcoholYes          1.313900   0.742487   1.770  0.07679 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120.285  on 89  degrees of freedom
## Residual deviance:  89.738  on 75  degrees of freedom
## AIC: 119.74
##
## Number of Fisher Scoring iterations: 5
```

```r
print("==== Intervention group ====")
```

```
## [1] "==== Intervention group ===="
```

```r
glm_wgt_chng <- glm(Losing_Wgt ~ Gender + age + Site + height + wgt_base + month +
    comorbid + Education + Marital + Tobacco + Alcohol, family = binomial, data = included.data[included
    "1", ])
summary(glm_wgt_chng)
```

```
##
## Call:
## glm(formula = Losing_Wgt ~ Gender + age + Site + height + wgt_base +
##     month + comorbid + Education + Marital + Tobacco + Alcohol,
##     family = binomial, data = included.data[included.data$tx ==
##         "1", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1315  -0.7559   0.4353   0.7927   1.5968
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           22.49627   10.51871   2.139   0.0325 *
## GenderFemale          -1.54250    1.27224  -1.212   0.2253
## age                    0.02918    0.03431   0.851   0.3950
## SiteSite 2            -0.59999    1.00023  -0.600   0.5486
## SiteSite 3            -1.81523    1.15333  -1.574   0.1155
## SiteSite 4            16.93634 2106.56415   0.008   0.9936
## SiteSite 5            -1.36191    1.09108  -1.248   0.2119
## height                -0.13118    0.05582  -2.350   0.0188 *
## wgt_base               0.03626    0.02368   1.532   0.1256
## month                 -0.22176    0.24209  -0.916   0.3597
## comorbid               0.07169    0.36631   0.196   0.8448
## Education> 12 years   -0.54453    0.79586  -0.684   0.4938
## MaritalMarried        -0.02912    0.66357  -0.044   0.9650
## TobaccoYes            -1.52045    0.90254  -1.685   0.0921 .
## AlcoholYes             1.20359    0.75429   1.596   0.1106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 93.188  on 75  degrees of freedom
## Residual deviance: 71.521  on 61  degrees of freedom
## AIC: 101.52
##
## Number of Fisher Scoring iterations: 17
```

The full model for CG can only find the significant effect of current tobacco use to losing weight, and the full model for IG shows the significant effect of height to losing weight. Now we perform model selections using a backward selection method. The best selected/reduced models for both groups are shown below.

```r
print("==== Control group ====")
```

```
## [1] "==== Control group ===="
```

```
glm_wgt_chng <- glm(Losing_Wgt ~ Gender + Tobacco + Alcohol, family = binomial, data = included.data[in
    "0", ])
summary(glm_wgt_chng)
```

```
##
## Call:
## glm(formula = Losing_Wgt ~ Gender + Tobacco + Alcohol, family = binomial,
##     data = included.data[included.data$tx == "0", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2625  -1.2002   0.5799   1.1547   1.4867
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.3225     0.9691  -2.397   0.0165 *
## GenderFemale   2.3760     0.9419   2.523   0.0116 *
## TobaccoYes     2.4254     0.9908   2.448   0.0144 *
## AlcoholYes     1.6197     0.6358   2.547   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120.28  on 89  degrees of freedom
## Residual deviance: 104.32  on 86  degrees of freedom
## AIC: 112.32
##
## Number of Fisher Scoring iterations: 5
```

```
print("==== Intervention group ====")
```

```
## [1] "==== Intervention group ===="
```

```
glm_wgt_chng <- glm(Losing_Wgt ~ Gender + height + wgt_base + Tobacco, family = binomial,
    data = included.data[included.data$tx == "1", ])
summary(glm_wgt_chng)
```

```
##
## Call:
## glm(formula = Losing_Wgt ~ Gender + height + wgt_base + Tobacco,
##     family = binomial, data = included.data[included.data$tx ==
##         "1", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8114  -1.2800   0.6524   0.8400   1.4282
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.08137    7.37469   2.045   0.0409 *
```

```
## GenderFemale -1.49878    0.96689  -1.550   0.1211
## height       -0.08663    0.04368  -1.983   0.0473 *
## wgt_base      0.01514    0.01480   1.023   0.3063
## TobaccoYes   -1.08050    0.67164  -1.609   0.1077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 93.188  on 75  degrees of freedom
## Residual deviance: 86.776  on 71  degrees of freedom
## AIC: 96.776
##
## Number of Fisher Scoring iterations: 4
```

The reduced model for CG shows that gender, tobacco use, and alcohol use are factors with significant effect to losing weight, while the reduced model for IG still can only find the significant effect of height to losing weight.

# Future Study Planning

## Imagine a further study that extends the questions of interest from your current study

A further study can investigate in the proportion of achieving outcome (losing weight) after the moderate behavioral weight loss program among tobacco users and non-smoking controls.

## Sample size for the new study

Goal: to achieve 80% power to detect 10% difference in proportion of losing weight after intervention among smoking group and non-smoking group. We would expect a higher percentage of losing weight in smoking group than non-smoking group. Assuming 70% of participants of the program losing weight. From current study, about 82.5% of participants were using tobacco, and 17.5% were not. Sample size for comparing two proportions:

```r
r = 82.5/17.5
(n <- ((r + 1)/r) * 0.7 * (1 - 0.7) * (0.84 + 1.96)^2/(0.1^2))
```

```
## [1] 199.5636
```

We would need 200 smokers and 200 non-smokers (400 at total) to achieve 80% power.

# Conclusion and Discussion

This study investigated the effectiveness of a moderate behavioral weight loss program for overweight adults, in comparison with a lower-intensity weight loss program as control group. There were 166 participants completed their intervention program and included in data analysis (n=76 in IG and n=90 in CG). The important findings are: First, participants in IG showed a significant weight loss (mean = -2.48 kg), and this amount was significantly greater than that in CG (mean = -0.79 kg, p-value=0.009), however the percentages of people losing weight after intervention between 2 groups were not different (p=61.1% in CG and p=69.7% in IG, p-value = 0.318); Second, two-sample Chi-squared test showed that proportion of individuals losing weight varied between clinical sites, specifically site 2 and site 4 observed the most and the least percentage (p=81.8% and p=41.7%, respectively); Third, the length of intervention program and race/ethnicity do not associate with the amount of weight loss as hypothesized, while only intervention condition-gender interaction and intervention condition-clinical site interaction had significant association to weight loss; Finally, gender and current tobacco and alcohol use are predictors for losing weight in low-intensity program, while losing weight status in moderate intervention can be significantly predicted by only height.

The strengths of this randomized controlled study are the abilities to establish the causal effect of intervention program, reduce bias, control for confounding, and determine treatment-predictive variable interactions. The weakness of this study design can be problem with generalisability (for example, the proportion of race/ethnicity in recruited sample might not be representative of the real-life distribution).

The results from this study have highlighted the effectiveness of the moderate weight loss program (mean weight loss of 2.48 kg over 11.7 months on average), and its superiority over the low-intensity program. This finding can encourage the implementation of this type of intervention for weight loss purpose. Future study can further investigate in the association between intervention time length from randomization and weight loss outcome by, for instance, recording participants' weight at several stages of time (such as every 3 months).