Final project report

Zhiwen (Owen) Jiang

2023-10-24

```
library(ggplot2)
```

Introduction to the data

An observational study was conducted to analyze the effect of smoking cessation on weight gain. The raw data contains 1629 subjects and 63 variables. The outcome is weight gain, which is the difference from baseline in weight (in kilograms), and the main factor is smoking cessation. Other variables include age, alcohol consumption, cholesterol level, diabetes, income, marital status, race, sex, etc.

Exploratory analysis

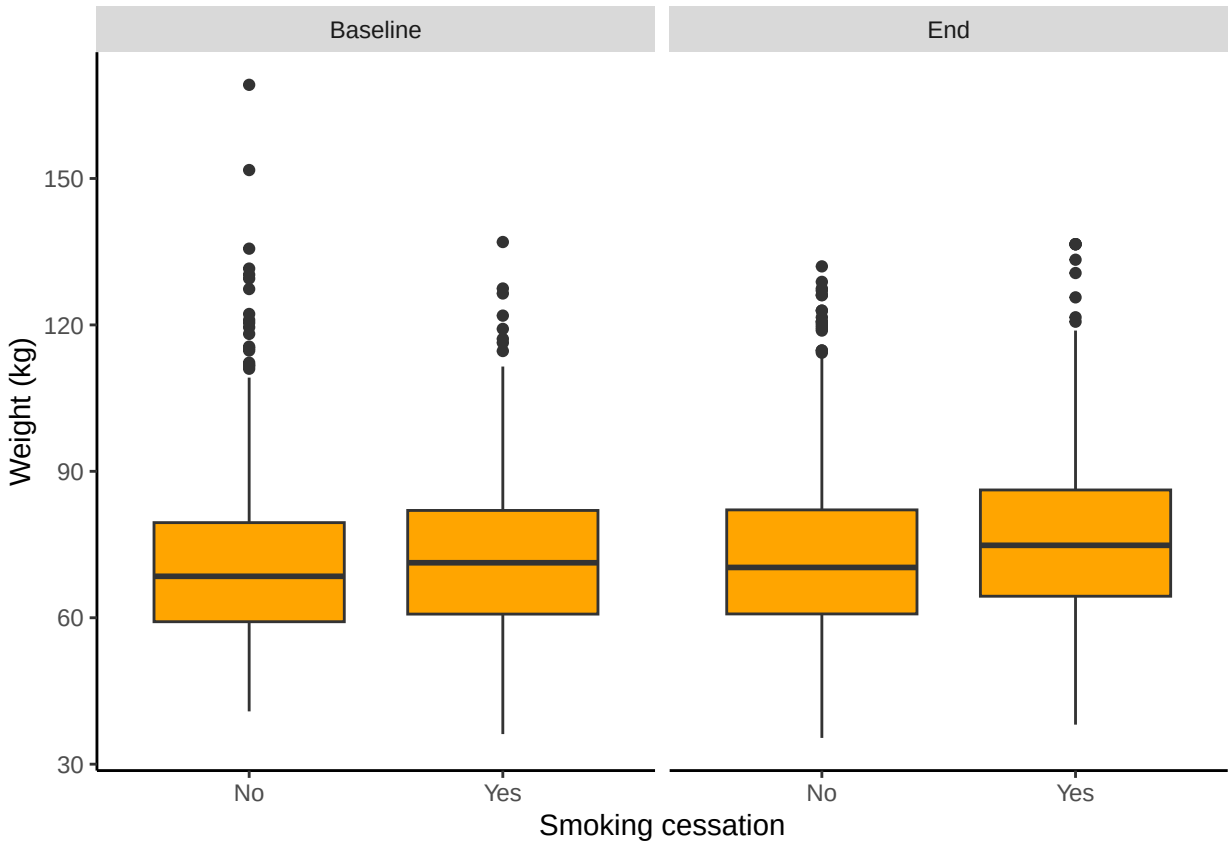
```
smk_raw <- read.csv('smoking_cessation.csv')
```

At the very beginning, we may want to know the weight distribution of smokers and non-smokers at baseline and at the end of study, respectively. We observe that people who quit smoking have higher weight at baseline and at the end of study. And people who quit smoking have a higher increment in weight.

```
plot1_data <- smk_raw[, c('seqn', 'wt_bl', 'wt_end', 'qsmk')]
plot1_data$qsmk <- factor(plot1_data$qsmk, levels = c(0, 1), labels = c('No', 'Yes'))
plot1_data <- reshape(plot1_data, direction = 'long', idvar = 'seqn',
                      varying = c(2, 3), sep='_')
plot1_data$time <- factor(plot1_data$time, levels = c('bl', 'end'),
                          labels = c('Baseline', 'End'))

ggplot(plot1_data, aes(y = wt, x = qsmk)) + geom_boxplot(fill = 'orange') + theme(panel.grid.major = el
        panel.background = element_blank(), axis.line = element_line(colour = "black")) + xlab('Smoking
```

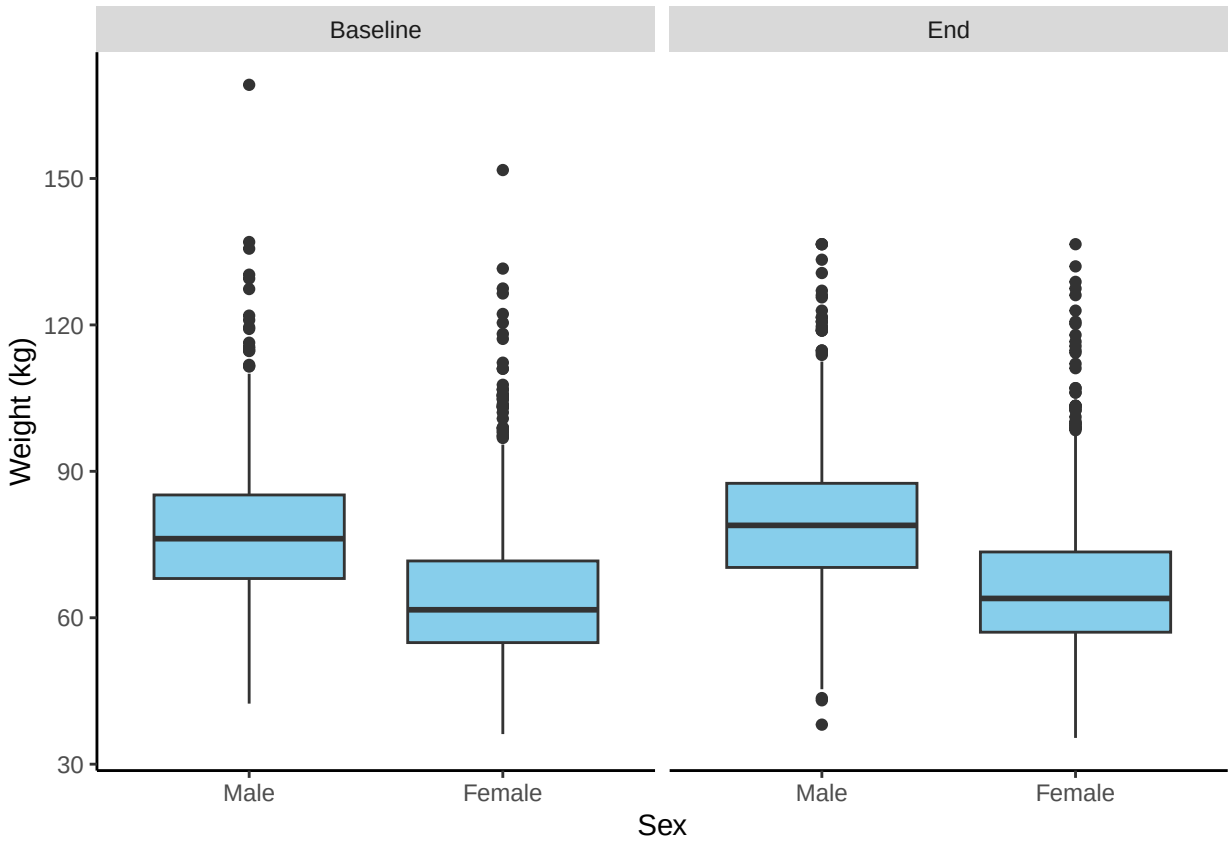
```
## Warning: Removed 63 rows containing non-finite values (`stat_boxplot()`).
```



Since sex is an important confounding for weight change, we also compare the difference. We cannot observe apparent difference of weight change between sexes.

```
plot2_data <- smk_raw[, c('seqn', 'wt_bl', 'wt_end', 'sex')]
plot2_data$sex <- factor(plot2_data$sex, levels = c(0, 1), labels = c('Male', 'Female'))
plot2_data <- reshape(plot2_data, direction = 'long', idvar = 'seqn',
                      varying = c(2, 3), sep='_')
plot2_data$time <- factor(plot2_data$time, levels = c('bl', 'end'),
                          labels = c('Baseline', 'End'))
ggplot(plot2_data, aes(y = wt, x = sex)) + geom_boxplot(fill = 'skyblue') + theme(panel.grid.major = el
        panel.background = element_blank(), axis.line = element_line(colour = "black")) + xlab('Sex') +
```

```
## Warning: Removed 63 rows containing non-finite values (`stat_boxplot()`).
```



Analysis for reason of quitting smoking.

In this section, we explore the reason that people quit smoking by fitting a logistic model. We include sex, age, race, marital status, educational level, weight at baseline, smoking intensity, smoking years and frequency of drinking alcohol as potential factors. Sex, age, race, smoking intensity and smoking years are statistically significant factors, where stronger smoking intensity keeps decreasing the probability of quitting smoking, while smoking years decrease the probability of quitting smoking at the beginning but increase it later.

```
lg_model <- glm(qsmk ~ sex + age + race + marital + education + wt_bl + smokeintensity + smokeyrs + alc
summary(lg_model)
```

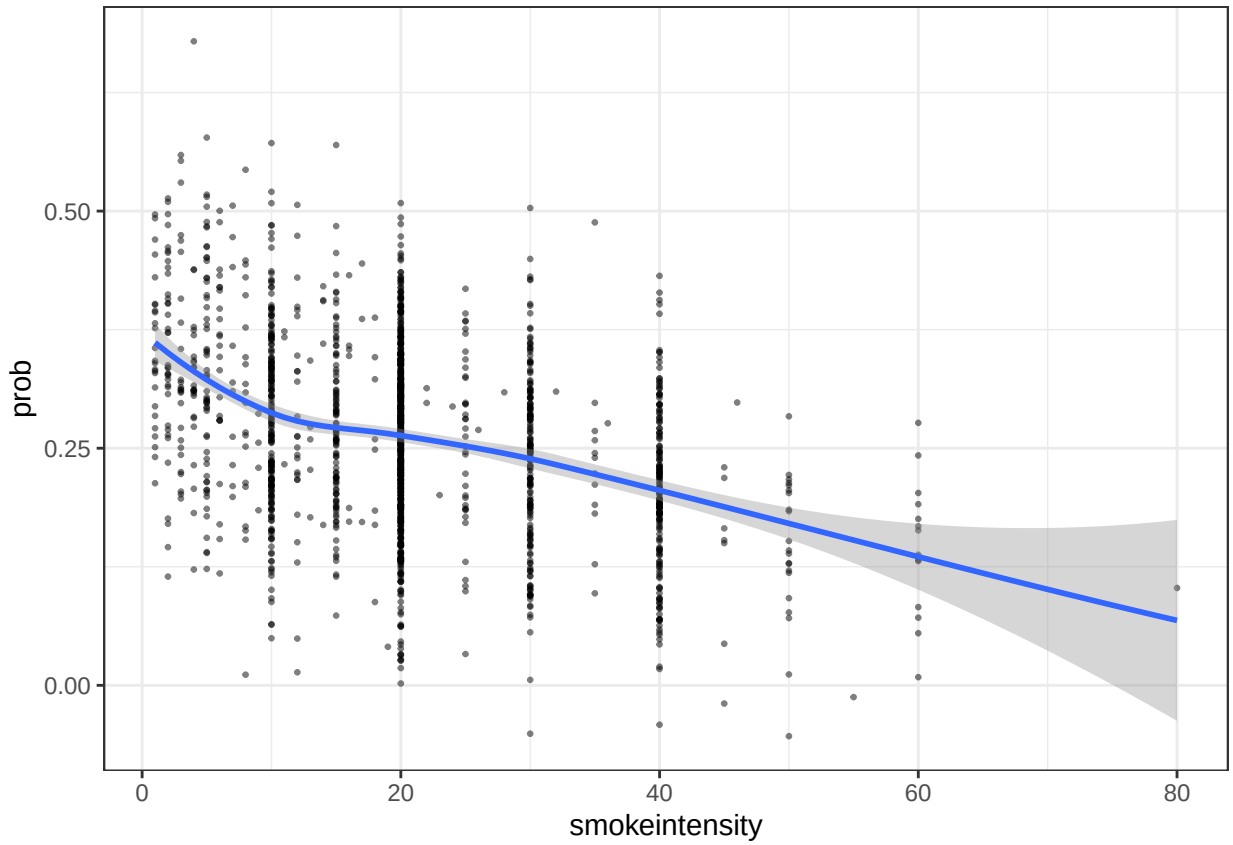
```
##
## Call:
## glm(formula = qsmk ~ sex + age + race + marital + education +
##     wt_bl + smokeintensity + smokeyrs + alcoholfreq, data = smk_raw)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5717  -0.2917  -0.2029   0.5151   0.9942
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.0222543  0.0910876  -0.244 0.807016
## sex            -0.0877730  0.0253705  -3.460 0.000555 ***
## age             0.0097692  0.0018434   5.299 1.32e-07 ***
## race           -0.1210767  0.0334797  -3.616 0.000308 ***
## marital        -0.0018441  0.0101044  -0.183 0.855208
## education       0.0170326  0.0097828   1.741 0.081860 .
```

```
## wt_bl           0.0012385  0.0007475   1.657 0.097722 .
## smokeintensity -0.0042680  0.0009584  -4.453 9.04e-06 ***
## smokeyrs       -0.0055766  0.0018878  -2.954 0.003182 **
## alcoholfreq     0.0071097  0.0086323   0.824 0.410279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1846454)
##
##     Null deviance: 315.55  on 1628  degrees of freedom
## Residual deviance: 298.94  on 1619  degrees of freedom
## AIC: 1883
##
## Number of Fisher Scoring iterations: 2
```

```
plot3_data <- smk_raw[, c('smokeintensity', 'smokeyrs')]
plot3_data$prob <- lg_model$fitted.values

ggplot(plot3_data, aes(x = smokeintensity, y = prob)) + geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(plot3_data, aes(x = smokeyrs, y = prob)) + geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```